異質なゲノムデータの相関解析法の開発とタンパク質ネットワーク予測への応用(Dissertation_全文)

AUTHOR(S):

山西, 芳裕

異質なゲノムデータの相関解析法の開発と タンパク質ネットワーク予測への応用

山西 芳裕

2005年

異質なゲノムデータの相関解析法の開発と タンパク質ネットワーク予測への応用

山西 芳裕

2005年

概要

ゲノム研究およびプロテオーム研究の進展とともに，遺伝子やタンパク質に関する大量の配列情報だけでなく，発現情報，進化情報，相互作用情報などがゲノムワイドに得られるようになった．ゲノム情報から生命システムへの情報構築原理を明らかにするためには，これらの個々のデータ解析はもちろん，データ間の比較や統合を行い，新しい生物学的な知見を得るための解析が望まれる．本研究では，これらの異質なゲノムデータを統一的に扱い，様々な解析をするための方法論の開発を行なった．

まず，ゲノムデータの一つである系統プロファイルの解析を行なった．ここで定義する系統プロファイルとは，ある遺伝子がゲノムの中に存在するかどうかを生物種ごとに0，1で表した文字列であり，様々な生物種に対してオーソログ遺伝子の有無を表した情報である．本研究では，独立成分分析という統計手法を用いた系統プロファイルの解析法を提案し，遺伝子の得失パターンに基づく生物種の分類を行なった．2875個のオーソログ遺伝子と77生物種から構成された系統プロファイルの解析の結果，9つの主要な生物種グループを抽出することができた．そして，各生物種グループに特徴的な遺伝子群を同定し，生物的機能との関係を検証した．また，生物種グループ間の階層性をクラスター分析によって検証したところ，古細菌は真正細菌よりも真核生物に近いという結果を示した．

次に，異質なゲノムデータの相関解析を行なった．同じ遺伝子またはタンパク質に関して，配列情報や，遺伝子発現，パスウェイ（タンパク質ネットワーク）などのデータが得られたときに，これらの異なる生物学的属性間の相関を解析することは重要であるが，データ構造が，文字列，数値ベクトル，グラフとそれぞれ異なるという問題がある．本研究では，カーネル正準相関分析という手法を提案することによって，複数の異質なゲノムデータ間の相関を解析し，その相関に寄与する遺伝子群を抽出する手法を開発した．実際に，提案した手法を大腸菌 *Escherichia coli K-12* のオペロン構造の検出法として適用した．代謝パスウェイ上の機能的な遺伝子産物間の関係，染色体上での遺伝子の隣接関係，マイクロアレイ実験で共発現する遺伝子間の関係を表す，パスウェイ，ゲノム，発現データの3つのデータを用いて，オペロンに属していると考えられる遺伝子群を探索的に抽出した．大腸菌のオペロンデータベースと比較し，予測精度を検証した結果，抽出した遺伝子群は，既知のオペロンに属する遺伝子群に，よく対応していることが確認できた．

最後に，本研究では，様々なゲノム情報から，生命システムを表すタンパク質ネットワークを予測する手法を開発した．異質なデータ間の相関解析を可能にし

たカーネル正準相関分析を用いて，ゲノムデータとタンパク質ネットワークの相関モデルを構築し，新規のタンパク質間ネットワークを予測する方法を提案した．この方法の独自性は，教師付き学習の枠組においてネットワーク推定を行なう点にある．実際の適用例として，出芽酵母 *Saccharomyces cerevisiae* のタンパク質間の機能ネットワークを，マイクロアレイ遺伝子発現情報，酵母2ハイブリッドシステムによる相互作用情報，タンパク質の細胞内局在情報，系統プロファイルの4種類のデータから予測した．実験によって判明している既知のタンパク質ネットワークを用いて評価した結果，本研究で提案する複数のデータの統合と教師付き学習の効果によって，先行研究の方法（教師なし学習）よりも予測精度が著しく向上することが確認できた．そこで，全てのタンパク質セットに対して提案手法を適用し，出芽酵母の6059個のタンパク質からなる機能的ネットワークを推定した．それを基に，未知のタンパク質の機能や，missing 酵素の遺伝子候補を予測し，その妥当性について検討し，この手法が新しい生物学的な発見に繋がる可能性について議論した．もう一つの適用例として，緑膿菌 *Pseudomonas aeruginosa* のリジン分解系におけるタンパク質ネットワークの再構築を試みた．ここでは，染色体上での遺伝子間の近さ，系統プロファイルによるタンパク質間の進化的な類似度を用いて，タンパク質の機能ネットワークを推定し，リジン分解系のパスウェイ上にあると思われる酵素遺伝子を予測した．EC:1.2.1.20，EC:2.6.1.48 などに対応すると予測された遺伝子について，大腸菌を宿主とした発現系を構築し酵素活性を確認したところ，実際に活性を示し，予測結果の妥当性を示唆した．

目 次

図目次

表目次

第1章　全体への序論

1.1　ポストゲノム時代におけるバイオインフォマティクス

ヒトをはじめ数多くの生物種において，ゲノムの全塩基配列（A, C, G, T の並び）が続々と決定されており，ゲノム配列が解読された生物種の数は増加の一途をたどっている．ゲノムの全塩基配列が決定されると，コード領域予測法，ホモロジー検索，モチーフ検索など，様々な計算法や統計解析法を駆使して，ゲノムにコードされている遺伝子産物のカタログが作成される．

しかしながら，これだけではゲノムの情報を真の意味で解読すること，すなわちゲノムから生命のはたらきを解読することはできない．配列の中に蓄えられる遺伝情報を明らかにし，その情報に基づいて組み立てられるタンパク質の立体構造や，生物の中でタンパク質が担っている機能とその構造との結びつきを解明していく必要がある．

これまでに機能が分かっている遺伝子との配列類似性が検出されない機能未知の遺伝子は，ゲノム配列解読後の各生物種において，3分の1から半数を占める．また，機能が推定された遺伝子の多くについても，その遺伝子産物同士の相互作用，つまり細胞内の複雑な分子間ネットワーク情報は分からない．それらを明らかにするためにも，配列情報だけでなく，これまでに蓄積された生命科学の膨大な知識はもちろん，DNA マイクロアレイなどの新しい実験データを活用し，生命システムを理解していく必要がある．

生物学的な機能とは，たくさんのタンパク質が互いに協調して働くことによって発揮されるものである．その意味で，生命のはたらきとは個々の遺伝子やタンパク質に帰せられるものではなく，多数の遺伝子あるいはタンパク質が複雑に相互作用したネットワークによって担われている．ゲノム情報から生命システムの構築原理の解明という目標に向けて，ゲノム情報を表す様々なデータを解析するための情報技術や統計手法を開発することが，ポストゲノム時代におけるバイオインフォマティクスの一つの役割と言えるであろう．

1.2 様々なゲノムデータや実験データの解析の必要性

ゲノム情報から生命システムを理解するという最終的な目的に向けて，ゲノムの配列情報はもちろん，遺伝子やタンパク質に関する様々な実験データを使って，遺伝子やタンパク質の機能やネットワークを明らかにしようという研究が，近年盛んになってきている．そのためのゲノム情報や実験データとして注目されているのが，DNA マイクロアレイによる遺伝子発現データ [14, 49]，酵母２ハイブリッドシステムによるタンパク質間相互作用のデータ [53, 26]，タンパク質の局在情報 [22]，系統プロファイル [43] などである．

DNA マイクロアレイを利用すれば，発生の様々な段階や異なる組織における細胞の遺伝子発現パターン，経時的な遺伝子発現の変化を系統的に調べることができる [14, 49]．同じような発現パターンを持つ遺伝子群は，同じような機能をもつであろうという仮定に基づき，クラスター分析や判別分析をすることによって，未知の遺伝子の機能を予測しようとする研究が盛んである [14, 10]．

酵母２ハイブリッドシステムは，転写因子のドメイン構造を巧みに利用し，タンパク質間の物理的な相互作用を検出する方法である [53, 26]．相互作用するタンパク質ペアは関連する機能を持つだろうという仮定に基づき，機能既知のタンパク質と機能未知のタンパク質の相互作用から，未知のタンパク質の機能を予測しようとする研究が近年盛んである．ただ，このデータは，ノイズが多く，疑陽性の相互作用が検出され易いという問題点も指摘されている．

タンパク質の細胞内局在情報に関しては，出芽酵母の全タンパク質の局在情報を網羅的に調べたデータが，近年，発表された [22]．GFP(Green Fluorescent Protein) で目的タンパク質をラベルすることにより，ゴルジ体，細胞質，小胞体，核内などの 23 個の細胞内局在のうち，出芽酵母のタンパク質が，どこで働いているかという情報を得ることができる．実験結果のデータは全て，データベースで公開されている [71]．

系統プロファイルとは，遺伝子がゲノムの中に存在するかどうかを生物種ごとに 0，1 で表した文字列であり，各オーソログ遺伝子を様々な生物種が持つかどうかを表した情報と解釈することができる [43]．先行研究では，各遺伝子のプロファイルは生物種毎の保存度を表すことから，それを一種の進化のパターンと考え，同じような系統プロファイルを持つ遺伝子ペアは，共進化の観点から同じような機能を持っていると仮定して，未知の遺伝子の機能予測を行う方法が提案されている [43, 55]．一方で，生物種間の比較という観点から見ると，各生物種のプロファイルは，進化の過程において遺伝子の得失を表した情報と解釈することができるので，これを用いたゲノムワイドな情報に基づく生物種の分類も可能である [32, 50]．しかしながら，系統プロファイルの研究自体が始まったばかりであり，実際に生

物種の分類や，タンパク質の機能予測に対して，どこまで有用性があるかは未知数の段階であり，さらなる研究が望まれている．

これらの様々なデータから生物学的に意味のある知見を得るためには，これらの大量データを効率良く解析する必要がある．また，単一のデータ解析だけでなく，異質なゲノムデータ間の相関解析，つまり異なる生物学的属性間の相関を解析する必要がある．例えば，遺伝子配列と遺伝子発現との関係，タンパク質の相互作用と遺伝子の進化との関係などの理解ができれば，生命現象の解明につながることが期待される．そういった生物学的な相関を探索的に検出するための手法が望まれるが，データ構造が，例えば，パスウェイはグラフ，配列は文字列，発現データは数値ベクトルなどとそれぞれ異なるので，複数のデータを同時に扱うのが困難という問題がある．実際にそれを実現するためには，文字列，数値ベクトル，グラフなど，データ構造が異なるデータを何らかの統一的な枠組で扱えるような方法論を開発する必要性がある．しかしながら，複数のデータを統合したり，同時に解析できるような手法は，ほとんど開発されていないのが現状である．

1.3 遺伝子やタンパク質間のネットワークの予測

ゲノム研究やプロテオーム研究の進展とともに，遺伝子及びタンパク質を中心とした生体分子間相互作用に関するゲノムワイドな実験データが蓄積されてきた．これらのゲノムワイドなデータから細胞レベルの網羅的な相互作用（ネットワーク）を理論的に推定することは，近年のバイオインフォマティクスにおける重要な研究テーマの一つである．例えば，遺伝子の転写制御に関するネットワーク，タンパク質間の物理的接触，代謝パスウェイなどが，ネットワーク推定における対象である．

実際に，遺伝子間や，タンパク質間の機能的な相互作用（ネットワーク）を，実験データから理論的に推定する方法を開発する研究が，盛んに行なわれるようになってきた．遺伝子制御ネットワークに対しては，マイクロアレイ実験から得られた遺伝子発現パターンから，ベイジアンネットワーク [17]，ブーリアンネットワーク [2]，微分方程式系 [12]，グラフィカルモデリング [51] などの手法を使って，遺伝子間の制御関係を予測する研究が盛んである．これらは，遺伝子発現パターンの類似度，時間的なずれ，または条件付独立などの統計的な尺度に基づき，相互作用する遺伝子ペアを選ぶアルゴリズムがほとんどである．タンパク質間相互作用ネットワークに対しては，例えば，タンパク質間の物理的な接触を予測する手法として，系統プロファイル法 [43] や，ミラーツリー法 [41]，インシリコ2ハイブリッド法 [42] などが提案されている．これらは，同じような進化のパターンを持つタンパク質ペアは，相互作用しやすいだろうという仮定に基づく．また，ゲノ

ム上の遺伝子の並びの特徴を利用した方法として，ゲノム上での遺伝子間の距離の近さ [8] や遺伝子融合の組合せ [15] に基づく方法などもある．複数のゲノム情報をグラフで表し，それを結合することによって，より信頼性のあるタンパク質間の機能的な関連を予測するジョイント法 [34] や，混合モデルのベイジアンネットワーク [27] などが提案されている．これまでに，上のような様々な数理的手法が提案されているが，どれも手法の理論的な提案だけに終わっており，未だ信頼できるネットワーク推定法は確立されていない．

実験科学と理論科学の共同研究が協調的に行われているゲノム配列解析と同様に，ネットワーク推定について理論解析の立場から実験科学への提言をおこない，新しい実験手法の開発の支援を行うためにも，様々な分子ネットワークを推定する信頼性のある手法の開発が望まれている．相互作用の予測は，新しい生物学的な知見に直接つながるため，学術的な期待はもちろん，ネットワーク推定の手法を産業界へ紹介することによる，新しい産業の発展への寄与が期待できる．

1.4 ゲノムデータ解析における統計学の役割

近年の生物科学は，大量の配列情報だけでなく，発現情報，進化情報，相互作用情報等，様々なデータを解析する必要に迫られている．それらの膨大なデータから効率的に新しい生物学的な知見を得るために，信頼性のある統計解析法の開発が望まれている．

塩基配列またはアミノ酸配列は，文字列のデータとして考えることができ，分子生物学において特徴的なデータである．配列比較やホモロジー検索では，Smith-Waterman アルゴリズム [48] による SSEARCH や，ヒューリスティックスを導入して高速計算可能な FASTA[60] や BLAST[3] などが開発されており，有効な手法として現在の分子生物学では不可欠のツールになっている．さらに，配列比較の結果，配列類似性の統計的有意性の評価に高度な数理統計の分布論が応用されている．

ゲノム配列からの遺伝子発見やモチーフ抽出などでは，隠れマルコフモデル [5, 31] の研究がさかんである．隠れマルコフモデルは統計学における一つのモデルであり，マルチプルアライメントや RNA2 次構造アライメント，モチーフ抽出などの配列解析に対して非常に強力なツールである．現在，隠れマルコフモデルを利用したタンパク質の機能予測システムや，モチーフのデータベースが構築されるなど，生物学的な知見を得るために，今では必要不可欠な統計解析法になっている．

近年，カーネル法 [46] と呼ばれる統計解析法が，バイオインフォマティクスの分野で注目を浴びるようになった．カーネル法は，機械学習や統計学の分野で開発された最先端の統計解析法であり，バイオインフォマティクスの分野では，特に，

サポートベクターマシン [9, 54] が有名である．サポートベクターマシンは，データの分類や判別を行う目的で開発された手法で，フィッシャーの判別分析や決定木などの既存の他の統計解析法よりも，はるかに高い分類性能を持つことが証明されている．特に，バイオインフォマティクスでは，遺伝子やタンパク質の分類や判別を目的として，頻繁に利用されるようになってきた．例えば，実際の適用例として，タンパク質のファミリー分類 [11]，タンパク質の細胞内局在予測 [39]，マイクロアレイ発現データからの遺伝子の機能予測 [10] や細胞や組織の分類 [18]，系統プロファイルからのタンパク質機能予測 [40, 55] 等が挙げられる．次の節で，簡単にカーネル法について紹介する．

1.5 カーネル法

ここでは，簡単にカーネル法のアイデアについて紹介する．カーネルとは，直感的には，オブジェクト間の類似度の尺度と考えることができる．本研究では，オブジェクトは遺伝子またはタンパク質に対応するので，カーネルとは，ゲノムデータが与えられたときの，遺伝子ペアまたはタンパク質ペアの類似度であると解釈することができる．

数学的には，2つのオブジェクト $\mathbf{x}$ と $\mathbf{x}'$ が与えられたとき，カーネルは，その特徴ベクトル $\phi(\mathbf{x})$ と $\phi(\mathbf{x}')$ の内積として次のように定義される．

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}').$$

つまり，$k(\mathbf{x}, \mathbf{x}')$ の値が大きければ大きいほど，$\mathbf{x}$ と $\mathbf{x}'$ は似ているということを意味し，$k(\mathbf{x}, \mathbf{x}')$ の値が小さければ小さいほど，$\mathbf{x}$ と $\mathbf{x}'$ は似ていないということを意味する．例として，ピアソンの相関係数なども一種のカーネルであると考えることができる．

仮に，n 個の遺伝子のセット $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ があるとしよう．各遺伝子の特徴を，例えば，A，C，G，T の4つの塩基の構成比で表すとすると，各遺伝子の特徴ベクトルは，

$$\begin{aligned} \phi(\mathbf{x}_1) &= (0.1, 0.4, 0.2, 0.3)^\top \\ \phi(\mathbf{x}_2) &= (0.2, 0.3, 0.3, 0.2)^\top \\ \vdots &\quad \vdots \\ \phi(\mathbf{x}_n) &= (0.5, 0.2, 0.1, 0.2)^\top, \end{aligned}$$

と表される．ここで，$\top$ は，ベクトルの転置を表す．このとき，遺伝子 $\mathbf{x}_1$ と遺伝子 $\mathbf{x}_2$ のカーネルは，それぞれの特徴ベクトルの内積をとって，

$$k(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2)$$

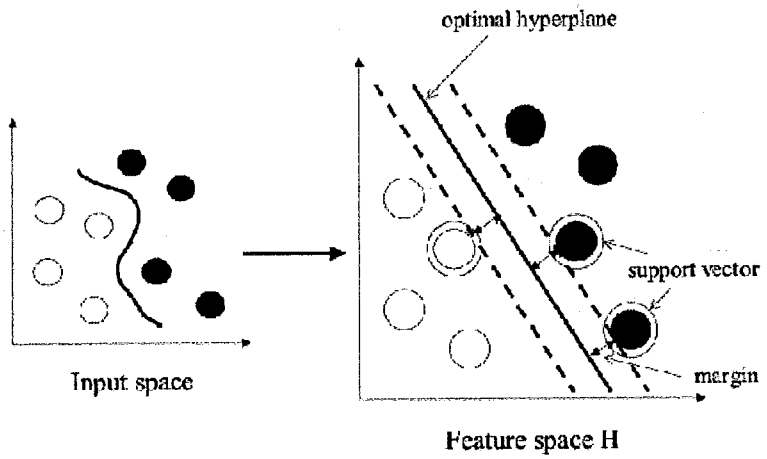

図 1.1: サポートベクターマシンによるデータの判別

$$
\begin{aligned}
&= 0.1 \times 0.2 + 0.4 \times 0.3 + 0.2 \times 0.3 + 0.3 \times 0.2 \\
&= 0.26,
\end{aligned}
$$

と計算される．この演算を全ての遺伝子ペアに対して実行したときに得られる類似度行列

$$
K = \begin{pmatrix}
k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\
k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_n) \\
\vdots & \vdots & \ddots & \vdots \\
k(\mathbf{x}_n, \mathbf{x}_1) & k(\mathbf{x}_n, \mathbf{x}_2) & \dots & k(\mathbf{x}_n, \mathbf{x}_n)
\end{pmatrix} = \begin{pmatrix}
0.30 & 0.26 & \dots & 0.21 \\
0.26 & 0.26 & \dots & 0.23 \\
\vdots & \vdots & \ddots & \vdots \\
0.21 & 0.23 & \dots & 0.34
\end{pmatrix},
$$

をカーネル行列と呼ぶ．

カーネル法とは，元のデータの入力空間 (Input space) ではなく，新しく類似度で構成された特徴空間 (Feature space) において，データの分類や特徴抽出などの様々な解析を行おうとするものである．特徴空間における遺伝子間の位置関係は，カーネル行列を反映したものとなる．

サポートベクターマシンを例にとって考えてみよう．図 1.1 は，そのアイデアを示している．白丸が正例，黒丸が負例のデータのとき，そのデータの判別問題を考える．元の入力空間では判別が難しいが， ある特徴空間に写像することによって，線形判別が可能となる．この判別ルールを，新しいデータが得られたときにも適用することによって，そのデータが正例に属するか負例に属するかを自動的に判別することができる．応用の例として，例えば，黒丸が癌細胞特異的に発現

することが分かっている遺伝子群，白丸が特異的に発現しないことが分かっている遺伝子群とすると，興味のある遺伝子が得られたとき，それが癌細胞特異的な遺伝子であるかどうか?の判別ができる．

カーネル法のアイデアを用いた手法として，サポートベクターマシン以外にも，カーネル主成分分析 [45]，カーネル正準相関分析 [1] などがあり，同様に近年のバイオインフォマティクスでよく利用されるようになってきた．なぜカーネル法がバイオインフォマティクスで注目されるかというと，カーネル法は，元のデータ構造に依存しないため，文字列，グラフ，数値ベクトルなど，どんな構造のデータに対しても，分類や予測，特徴抽出などが行えるという長所があるからだと考えられる．

1.6 本研究の目的と概要

近年の生物工学の進歩により，大量の配列情報だけでなく，発現情報，進化情報，相互作用情報等がゲノムワイドに得られるようになった．本研究では，個々のゲノムデータ解析だけでなく，異質なゲノムデータを統一的に扱い，様々な解析をするための方法論の開発を行なった．本論文は，主に3つのトピックで構成されている．系統プロファイルを用いた生物種グループの解析 (2章)，異質なゲノムデータの相関解析法の開発 (3章)，新規のタンパク質ネットワークの予測法の開発 (4章) であり，それらの概要を以下に示す．

まず，ゲノムデータの一つである系統プロファイルの解析を行い，生物種グループの抽出を行った [62]．ここで定義する系統プロファイルとは，ある遺伝子がゲノムの中に存在するかどうかを生物種ごとに0，1で表した文字列であり，様々な生物種に対してオーソログ遺伝子の有無を表した情報である．本研究では，独立成分分析という統計手法を用いた系統プロファイルの解析法を提案し，遺伝子の得失パターンに基づく生物種の分類を行なった．2875個のオーソログ遺伝子と77生物種から構成された系統プロファイルの解析の結果，9つの主要な生物グループを抽出することができた．そして，各生物グループに特徴的な遺伝子群を同定し，生物的機能との関係を検証した．また，生物グループ間の階層性をクラスター分析によって検証したところ，古細菌は真正細菌よりも真核生物に近いという結果を示した．

次に，異質なゲノムデータの相関解析を行う手法を開発し，様々なゲノムデータを用いてオペロンの解析を行った [63]．同じ遺伝子またはタンパク質に関して，配列情報や，遺伝子発現，パスウェイ（タンパク質ネットワーク）などのデータが得られたときに，これらの異なる生物学的属性間の相関を解析することは重要であるが，データ構造が，文字列，数値ベクトル，グラフとそれぞれ異なるとい

う問題がある．本研究では，カーネル正準相関分析という手法を提案することによって，複数の異質なゲノムデータ間の相関を解析し，その相関に寄与する遺伝子群を抽出する手法を開発した．実際に，提案した手法を大腸菌 *Escherichia coli K-12* のオペロン構造の検出法として適用した．代謝パスウェイ上の機能的な遺伝子産物間の関係，染色体上での遺伝子の隣接関係，マイクロアレイ実験で共発現する遺伝子間の関係を表す，パスウェイ，ゲノム，発現データの３つのデータを用いて，オペロンに属していると考えられる遺伝子群を探索的に抽出した．大腸菌のオペロンデータベースと比較し，予測精度を検証した結果，抽出した遺伝子群は，既知のオペロンに属する遺伝子群に，よく対応していることが確認できた．

最後に，様々なゲノム情報から，生命システムを表すタンパク質ネットワークを予測する手法を提案した [64]．異質なデータ間の相関解析を可能にしたカーネル正準相関分析を用いて，ゲノムデータとタンパク質ネットワークの相関モデルを構築し，新規のタンパク質間ネットワークを予測する方法を提案した．この方法の独自性は，教師付き学習の枠組においてネットワーク推定を行なう点にある．実際の適用例として，出芽酵母のタンパク質間の機能ネットワークを，マイクロアレイ遺伝子発現情報，酵母２ハイブリッドシステムによる相互作用情報，タンパク質の細胞内局在情報，系統プロファイルの４種類のデータから予測した．実験によって判明している既知のタンパク質ネットワークを用いて評価した結果，本研究で提案する複数のデータの統合と教師付き学習の効果によって，先行研究の方法（教師なし学習）よりも予測精度が著しく向上することが確認できた．そこで，全てのタンパク質セットに対して提案手法を適用し，出芽酵母 *Saccharomyces cerevisiae* の 6059 個のタンパク質からなる機能的ネットワークを推定した．それを基に，未知のタンパク質の機能や，missing 酵素の遺伝子候補を予測し，その妥当性について検討し，この手法が新しい生物学的な発見に繋がる可能性について議論した．もう一つの適用例として，緑膿菌 *Pseudomonas aeruginosa* のリジン分解系におけるタンパク質ネットワークの再構築を試みた．ここでは，染色体上での遺伝子間の近さ，系統プロファイルによるタンパク質間の進化的な類似度を用いて，タンパク質の機能ネットワークを推定し，リジン分解系のパスウェイ上にあると思われる酵素遺伝子を予測した．EC:1.2.1.20，EC:2.6.1.48 などに対応すると予測された遺伝子について，大腸菌を宿主とした発現系を構築し酵素活性を確認したところ，実際に活性を示し，予測結果の妥当性を示唆した．

第2章　系統プロファイルを用いた生物種グループの抽出

2.1　序論

様々な生物種のゲノム配列が解読されるにともない，系統プロファイル [43] に基づいたオーソログ遺伝子の解析がさかんに行われており，バイオインフォマティクスにおける新しいアプローチとして注目を浴びている．ここで定義する系統プロファイルとは，ある遺伝子またはタンパク質がゲノムの中にコードされているかどうかを生物種ごとに 0，1 で表した文字列であり，遺伝子の得失を表した情報と解釈することができる．あるオーソログ遺伝子を生物種が持っているかどうかの判定は，全配列ペアに対するホモロジー検索に基づく配列類似性の有無で評価される．

先行研究では，各遺伝子のプロファイルは生物種毎の遺伝子の保存度を表すことから，それを一種の進化のパターンと考え，共進化の観点から，未知の遺伝子の機能予測を行う方法が提案されている [43, 34, 55]．一方で，生物種間の比較という観点から見ると，各生物種のプロファイルは，進化の過程において遺伝子の得失を表した情報と解釈することができるので，これを用いたゲノムワイドな情報に基づく生物種の分類が考えられる [50, 32, 59]．伝統的な分子進化学の分野における先行研究として，特定の塩基配列やアミノ酸配列の多重アライメントを基にした系統分類法の研究は盛んに行われているが，ゲノムワイドな情報から得られた系統プロファイルに基づく系統分類法の研究はほとんど行われていない．

本研究では，独立成分分析 [13, 23, 24] という統計手法を用いた系統プロファイルの解析法を提案した [62]．遺伝子の獲得パターンという意味で，多数の生物種から同じような系統パターンを持っている生物をまとめあげ，小数の特徴的な系統パターンを表す生物種のグループを抽出するのが目的である．つまり，それは高次元の生物種で構成された空間から，生物種グループで構成される空間に射影することに相当する．図 2.1 は，本研究の目的の概略図を示す．

データとして，京都大学化学研究所の生命情報データベース KEGG(Kyoto Encyclopedia of Genes and Genomes)[29, 69] に登録されている，2875 個のオーソログ遺伝子，77 生物種から構成された系統プロファイルを用いた．提案手法によっ

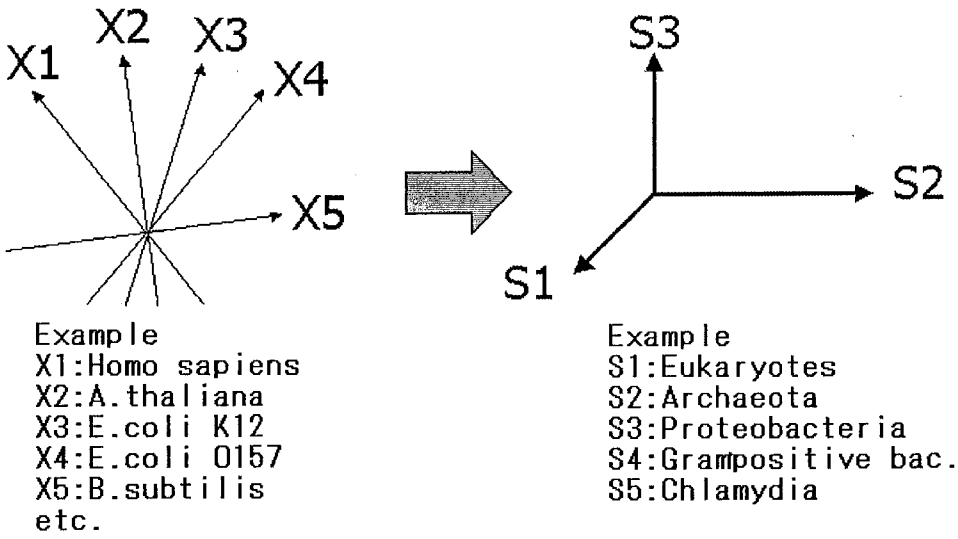

図 2.1: 生物種で構成される空間から，生物種グループで構成される空間への射影

て抽出された独立成分と主要な生物種グループの対応に注目し，独立成分と生物種との相関係数を基にした新しい系統分類法を開発した．生物種グループ間の階層性をクラスター分析によって検証したところ，古細菌は真正細菌よりも真核生物に近いという結果を示した．また，抽出された成分の高得点，低得点の遺伝子に注目し，各生物種グループに特徴的な遺伝子群と生物的機能との関係を検証した．

2.2 方法

2.2.1 系統プロファイルの構築

系統プロファイルは，2002年5月時点で，KEGGデータベース[28, 69]で定義されている2875個のオーソログ遺伝子を，77種の生物種が持っているか持っていないかを基準に作成した．ここでは，ゲノム配列が完全に解読できている生物種だけを用いている．その内訳は，真核生物は6種類，古細菌は13種類，真正細菌は58種類である．表2.1と表2.2は，この解析で用いた全生物種の名前とKEGGにおける省略名，各生物種が真核生物(Eukaryotes)，古細菌(Archaea)，真正細菌(Bacteria)の生物界の3つの主要ドメインのうち，どのドメインに属するかを示す．

各オーソログ遺伝子の系統プロファイルは，オーソログ遺伝子の有無にしたがって，上の77種の生物種に対して1と0で構成されるビット列である．表2.3は，

表 2.1: 生物種名と省略名のリストI

番号	省略名	生物種	主要ドメイン
1	hsa	Homo sapiens	Eukaryotes
2	dme	Drosophila melanogaster	Eukaryotes
3	cel	Caenorhabditis elegans	Eukaryotes
4	ath	Arabidopsis thaliana	Eukaryotes
5	sce	Saccharomyces cerevisiae	Eukaryotes
6	spo	Schizosaccharomyces pombe	Eukaryotes
7	mja	Methanococcus jannaschiin	Archaea
8	mth	Methanobacterium thermoautotrophicum	Archaea
9	afu	Archaeoglobus fulgidus	Archaea
10	mka	Methanopyrus kandleri	Archaea
11	tac	Thermoplasma acidophilum	Archaea
12	tvo	Thermoplasma volcanium	Archaea
13	pho	Pyrococcus horikoshii	Archaea
14	pab	Pyrococcus abyss	Archaea
15	pfu	Pyrococcus furiosus	Archaea
16	ape	Aeropyrum pernix	Archaea
17	sso	Sulfolobus solfataricus	Archaea
18	sto	Sulfolobus tokodaii	Archaea
19	pai	Pyrobaculum aerophilum	Archaea
20	eco	Escherichia coli K-12 MG1655	Bacteria
21	ecj	Escherichia coli K-12 W3110	Bacteria
22	ece	Escherichia coli O157 EDL933	Bacteria
23	ecs	Escherichia coli O157 Sakai	Bacteria
24	sty	Salmonella typhi	Bacteria
25	stm	Salmonella typhimurium	Bacteria
26	ype	Yersinia pestis	Bacteria
27	hin	Haemophilus influenzae	Bacteria
28	pmu	Pasteurella multocida	Bacteria
29	xfa	Xylella fastidiosa	Bacteria
30	vch	Vibrio cholerae	Bacteria
31	pae	Pseudomonas aeruginosa	Bacteria
32	buc	Buchnera sp. APS	Bacteria
33	nme	Neisseria meningitidis MC58 (serogroup B)	Bacteria
34	nma	Neisseria meningitidis Z2491 (serogroup A)	Bacteria
35	rso	Ralstonia solanacearum	Bacteria
36	hpy	Helicobacter pylori 26695	Bacteria
37	hpj	Helicobacter pylori J99	Bacteria
38	cje	Campylobacter jejuni	Bacteria
39	rpr	Rickettsia prowazekii	Bacteria
40	rco	Rickettsia conorii	Bacteria

系統プロファイルのデータセットの例を示したものである．本研究の場合，遺伝子の数$n = 2875$，生物種の数$p = 77$の系統プロファイルとなる．行ごとの視点から見ると，各オーソログ遺伝子は，77個のビット列のプロファイルとなる．列ごとの視点から見ると，各生物種は，2875個のビット列のプロファイルとなる．これは，進化の過程において遺伝子の得失を表した情報と解釈することができる．

2.2.2 独立成分分析

系統プロファイルは，サンプルを遺伝子，変数を生物種と見れば，一種の多変量データだと見なすことができる．一般的に，多変量データの主な概観を理解するために，少ない次元に特徴量を落として解釈をし易くしようと言うのは，データ解析において必要なことである．独立成分分析 (independent component analysis

表 2.2: 生物種名と省略名のリストII

番号	省略名	生物種	主要ドメイン
41	mlo	Mesorhizobium loti	Bacteria
42	sme	Sinorhizobium meliloti	Bacteria
43	atu	Agrobacterium tumefaciens C58 (UWash/Dupont)	Bacteria
44	atc	Agrobacterium tumefaciens C58 (Cereon)	Bacteria
45	bme	Brucella melitensis	Bacteria
46	ccr	Caulobacter crescentus	Bacteria
47	bsu	Bacillus subtilis	Bacteria
48	bha	Bacillus halodurans	Bacteria
49	sau	Staphylococcus aureus N315 (MRSA)	Bacteria
50	sav	Staphylococcus aureus Mu50 (VRSA)	Bacteria
51	lmo	Listeria monocytogenes	Bacteria
52	lin	Listeria innocua	Bacteria
53	lla	Lactococcus lactis	Bacteria
54	spy	Streptococcus pyogenes SF370 (serotype M1)	Bacteria
55	spn	Streptococcus pyogenes MGAS8232 (serotype M18)	Bacteria
56	spr	Streptococcus pneumoniae R6	Bacteria
57	cac	Clostridium acetobutylicum	Bacteria
58	cpe	Clostridium perfringens	Bacteria
59	mge	Mycoplasma genitalium	Bacteria
60	mpn	Mycoplasma pneumoniae	Bacteria
61	mpu	Mycoplasma pulmonis	Bacteria
62	uur	Ureaplasma urealyticum	Bacteria
63	mtu	Mycobacterium tuberculosis H37Rv (lab strain)	Bacteria
64	mtc	Mycobacterium tuberculosis CDC1551	Bacteria
65	mle	Mycobacterium leprae	Bacteria
66	ctr	Chlamydia trachomatis	Bacteria
67	cmu	Chlamydia muridarum	Bacteria
68	cpn	Chlamydophila pneumoniae CWL029	Bacteria
69	cpa	Chlamydophila pneumoniae AR39	Bacteria
70	cpj	Chlamydophila pneumoniae J138	Bacteria
71	bbu	Borrelia burgdorferi	Bacteria
72	tpa	Treponema pallidum	Bacteria
73	syn	Synechocystis sp. PCC6803	Bacteria
74	ana	Anabaena sp. PCC7120 (Nostoc sp. PCC7120)	Bacteria
75	dra	Deinococcus radiodurans	Bacteria
76	aae	Aquifex aeolicus	Bacteria
77	tma	Thermotoga maritima	Bacteria

(ICA)) は，統計学の分野で開発された特徴抽出の方法論である [13, 23, 24]．この手法の目的は，多変量データの中に潜んでいる興味深い潜在変量を抽出しようというもので，実際の応用では，特徴抽出はもちろん，音声認識，多変量データの次元縮小など，たくさんの用途で利用されている．独立成分分析は，バイオインフォマティクスの分野でも，遺伝子発現データの解析 [33] に用いられており，その有効性が示されているが，系統プロファイルなどの配列や進化のデータに対しての適用例はない．

簡単に，独立成分分析のアイデアについて説明する．観測変数ベクトルとして，$\mathbf{x} = (x_1, x_2, \cdots, x_p)^T$ が与えられたとする．ここで，p は，変数の数であり，独立成分分析の数学モデルは，以下のように定式化される．

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.1}$$

表 2.3: 系統プロファイル：n 個の遺伝子，p 個の生物種の例

	生物種 1	生物種 2	生物種 3	生物種 4	生物種 5	⋯	生物種 p
遺伝子 1	1	1	0	0	0	⋯	1
遺伝子 2	1	0	1	0	1	⋯	0
遺伝子 3	0	1	0	0	1	⋯	1
遺伝子 4	0	1	0	1	0	⋯	0
遺伝子 5	1	1	1	0	1	⋯	0
遺伝子 6	0	0	0	0	1	⋯	1
遺伝子 7	0	1	1	1	0	⋯	1
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
遺伝子 n	1	0	1	0	0	⋯	1

ここで，$\mathbf{s} = (s_1, s_2, \cdots, s_m)^T$ は，統計的に独立な潜在変量ベクトルであり，独立成分と呼ばれる．m は，独立成分の個数であり，$\mathbf{A}$ は，未知の何らかの混合行列である．潜在変数 $\mathbf{s}$ の実際の値は，独立成分スコアと呼ばれ，$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} = \mathbf{W}\mathbf{x}$ と計算される．独立成分 $\mathbf{s}$ を求めることは，実際，重み行列 $\mathbf{W}$ を求めることに相当する．各独立成分の分散は定義することができないため，ここでは，全て分散は 1 と定義し，$E\{s_j^2\} = 1$ とする．

独立成分分析に似た手法として主成分分析 (principal component analysis (PCA))[4] がある．主成分分析では，潜在変数が無相関で正規分布に従うことを仮定しているのに対し，独立成分分析では，潜在変数が独立で非正規分布に従うことを仮定している．独立性の条件は，数学的に無相関よりも強いため，ICA と PCA の実際の適用における性能は異なる．アルゴリズムの観点からは，主成分分析では，潜在変数の分散を最大にすることを考えるのに対して，独立性分分析では，潜在変数の非正規性を最大にすることを考える．これによって，実際の応用においては，単なる次元縮小を目的とする主成分分析よりも独立成分分析の方が，抽出した成分ごとに意味を持つので解釈が容易であることが知られている．

独立成分を $\mathbf{y} = \mathbf{W}\mathbf{x}$ と求めたいと仮定する．ここで，$\mathbf{y} = (y_1, y_2, \cdots, y_m)^T$ である．非正規性の一つの尺度としてネグエントロピーを考え，その最大化を考える．通常，エントロピーは，$\mathbf{y}$ の確率密度関数を $f(\mathbf{y})$ としたとき，以下のように定義できる．

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y}. \tag{2.2}$$

ここで，$\mathbf{y}$ のネグエントロピーは，

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}), \tag{2.3}$$

と定義できる．そのとき，$\mathbf{y}_{gauss}$ は，$\mathbf{y}$ の共分散行列に基づくランダムな正規分布に従う成分である．最終的に，独立成分の推定は，$J(\mathbf{y})$ を最大化するような $\mathbf{y}$ を求めることに帰着する．この計算を効率的に計算する FastICA[23, 24] と呼ばれる手法が開発されており，本研究ではこれを用いる．アルゴリズムの詳細は，参考文献 [23, 24] を参照されたい．

2.2.3 独立成分と生物種グループとの関連付け

変数 x は，系統プロファイルにおける生物種に相当するとし，変数 y を，独立成分に相当すると考える．独立成分が，ある特定の生物種グループに対応しているかどうかをみるために，次のようなピアソンの相関係数を用いることを考えた．

$$r(x,y) = \frac{1/n\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(1/n\sum_{i=1}^{n}x_i-\bar{x})^2}\sqrt{(1/n\sum_{i=1}^{n}y_i-\bar{y})^2}}, \tag{2.4}$$

ここで，n は，遺伝子の個数であり，$\bar{x}$ と $\bar{y}$ は，x と y の実際の観測値の平均である．つまり，この相関係数が高ければ，独立成分 y は，生物種 x が属する生物種グループであると判定することができる．逆に，この相関係数がゼロに近ければ，独立成分 y と生物種 x は，生物種グループとしては，あまり関連が無いと判定することができる．

2.3 結果

2.3.1 生物種グループの抽出

系統プロファイルのデータセットは，行が 2875 個の遺伝子，列が 77 種の生物種に相当する，2875 × 77 の多変量データと見なすことができる．ここで対象とする 77 種の生物種は，分子生物学的な事前知識から，多くても 18 グループ程度に分かれるのではないかと推測した．そこで，求めたい独立成分の個数を 18 と設定して独立成分分析を実行し，2875 × 77 の系統プロファイル行列を，2875 × 18 の独立成分行列に変換した．つまり，生物種グループの候補として，18 個の独立成分を抽出することができた．

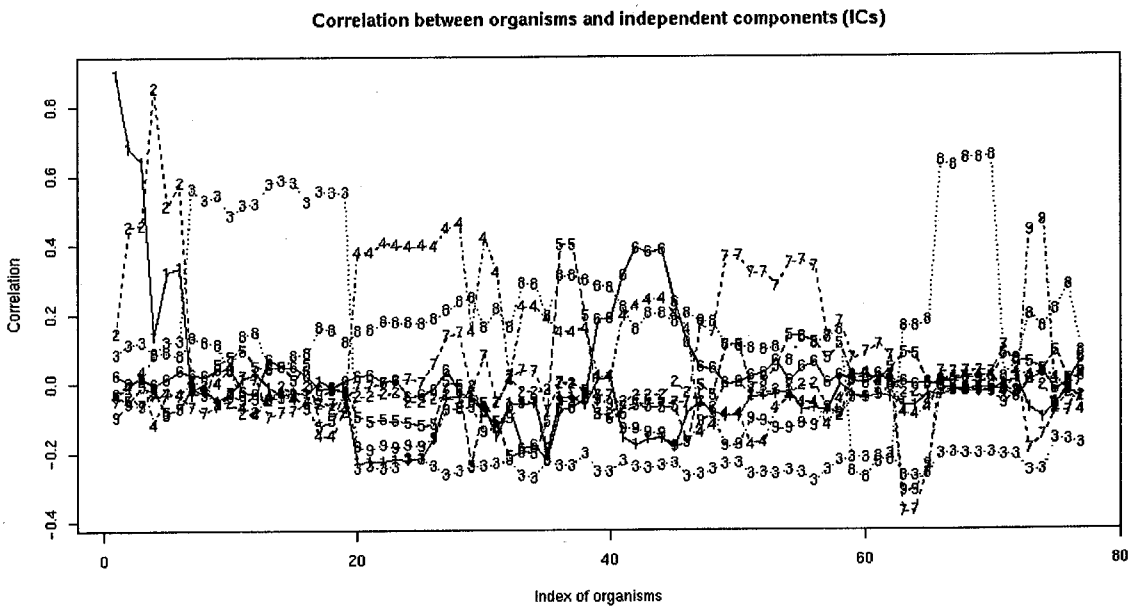

図 2.2: 生物種と独立成分との相関係数プロット：
横軸は，真核生物，古細菌，真正細菌の順で並ぶ，生物種のインデックスを示しており，縦軸は相関係数の強さを表す．グラフ中の 1 から 9 の番号は，独立成分の番号を示しており，相関係数の高いものを生物種の順に並べている．

2.3.2 生物種のグループ分け

抽出した独立成分を生物学的に解釈するために，18個の独立成分と，77個の生物種から系統ベクトルとの相関係数を，全ての組合せに対して計算した．その結果，18個の成分のうち9個が，ある特定の生物種グループに良く相関していることが分かった．図2.2は，その結果を示す．ここで，相関係数が，ある生物種のところでピークの山を作っていることが分かる．これは，独立成分が，ある生物種グループに対して正の寄与をしていると考えることができる．この相関係数のピークに相当した生物種が，KEGGデータベースで提供されている生物種グループにどのように対応するか調べた．表2.4は，9個の独立成分と生物種グループとの対応関係を要約している．77種の生物種のうち，*Deinococcus raiodurans*, *Aquifex aeolicus*, *Thermotoga maritima* 以外の74種までが，その9個の独立成分で説明されることが分かった．

表 2.4: 抽出した独立成分を生物種グループとして解釈した結果

独立成分	生物種グループ	種の例
IC1	Eukaryotes(animal)	*Homo sapiens, D. melanogaster*, etc.
IC2	Eukaryotes(plant/fungi)	*A. thaliana, S. cerevisiae*, etc.
IC3	Archaea	*M. jannaschii*, T. acidophilum, etc.
IC4	Proteobacteria(gamma)	*E. coli, S. typhi*, etc.
IC5	Proteobacteria(delta/epsilon)	*H. pylori, R. solanacearum*, etc.
IC6	Proteobacteria(alpha)	*M. loti, S. meliloti*, etc.
IC7	Grampositive bacteria (Low G+C)	*B. subtilis, B. halodurans*, etc.
IC8	Chlamydia	*C. trachomatis, C. muridarum*, etc.
IC9	Cyanobacteria	*Synechocystis sp., Anabaena sp.*

2.3.3 生物種の階層性

元の系統プロファイルのセットを列ごとに見ることによって，生物種の分類に使うことも可能である．ユークリッド距離(ここでは1と0のデータなのでハミング距離に相当する)を用いて，生物種間の非類似度を定義して，階層的クラスタリングを行った．ここでは，最長距離法を採用した．図2.3は，その結果を示しており，ここでラベルはKEGGデータベースにおける生物種名の省略形である．

次に，独立成分分析による解析結果に基づく，生物種のクラスター分析を実行

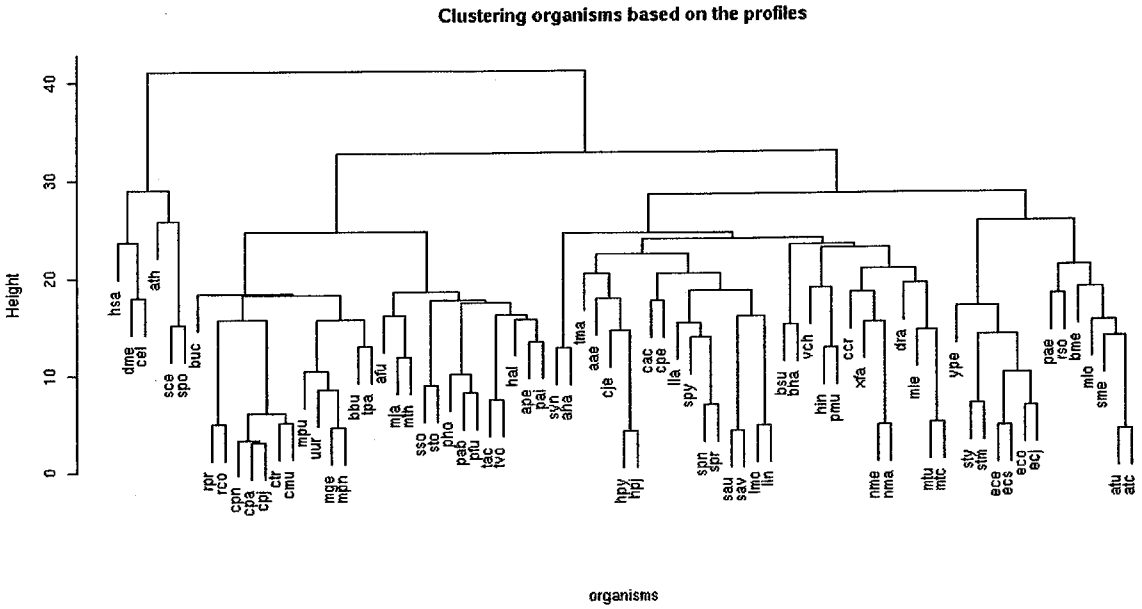

図 2.3: 元の系統プロファイルに基づく生物種のクラスタリング：
デンドログラムのラベルは，KEGG データベースにおける生物種名の省略名を表す．

した．図 2.2 で示した，生物種と独立成分との相関係数は，各生物種に対し，9 次元のベクトルのセットであると考えることができる．これらの相関ベクトルに対して，生物種間のユークリッド距離を計算し，クラスタリングを実行した結果が，図 2.4 である．比較の結果，後者の方が，KEGG の生物種のグループ分類に良く相当していることが分かった．(KEGG の生物種グループ分類は，NCBI[70] の分類に基づいている．) 18 個全ての独立成分を使って，同じクラスター分析を実行したところ，図 2.3 で示している前者の分類結果と同じような結果になった．

2.3.4 生物種グループ特異的遺伝子の抽出

独立成分分析の特長は，各生物種グループへの遺伝子の帰属度を，数量化できるところにある．各生物種グループ軸において，全遺伝子の分布を見るために，生物種グループに対応する独立成分スコアの散布図をプロットした．図 2.5 は，独立成分のスコアをプロットした図であり，それぞれ，第 1 独立成分と第 2 独立成分 (IC1 vs. IC2), 第 3 独立成分と第 4 独立成分 (IC3 vs. IC4), 第 5 独立成分と第 6 独

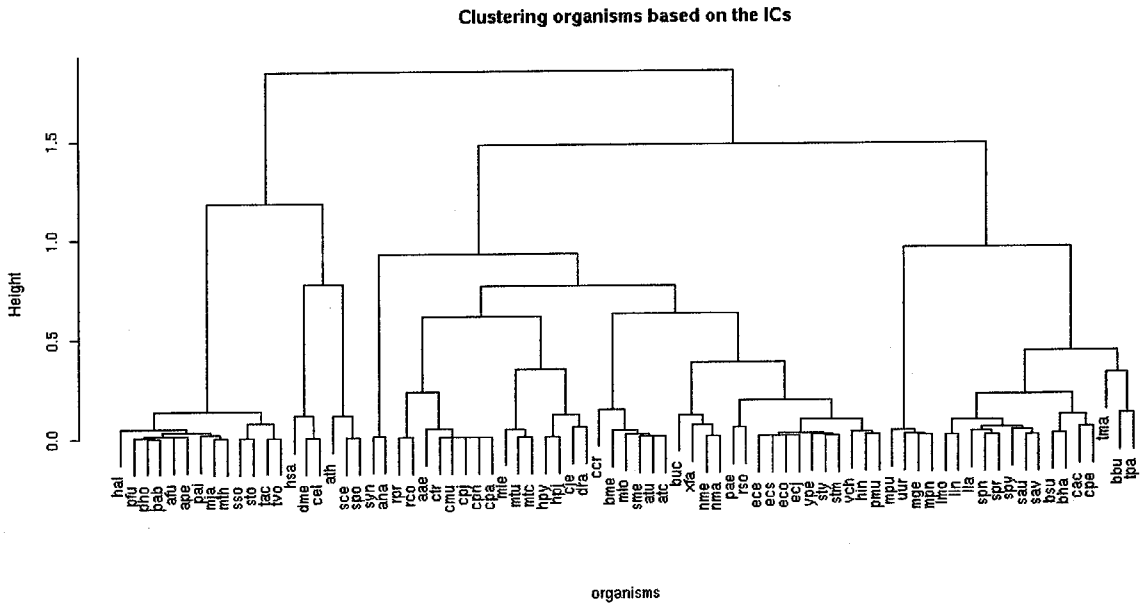

図 2.4: 生物種と独立成分間の相関係数ベクトルに基づく生物種のクラスタリング：デンドログラムのラベルは，KEGG データベースにおける生物種名の省略名を表す.

立成分 (IC5 vs. IC6), 第7独立成分と第8独立成分 (IC7 vs. IC8) の散布図をそれぞれ示している．これらのプロットにおいて，各独立成分で高得点の遺伝子群は，生物種特異的な遺伝子群であると解釈できる．

そこで，各生物種グループに対応する軸で，高得点の遺伝子群を抽出することで，ある生物種グループ特異的な遺伝子群の検出を試みた．表 2.5 と表 2.6 は，各独立成分で，5% 分位点以上の得点を持つ遺伝子の個数を数えた結果を示している．これを見ると，独立成分と生物種の間の関係が明らかである．例えば，第1独立成分は，動物のグループの遺伝子で占められていることが分かる．

生物学的な機能という観点から，抽出した独立成分の妥当性を検証するために，抽出した遺伝子が，KEGG のパスウェイにどのように載るかを調べた．一例として，第4独立成分で高得点の遺伝子群に注目した．この第4独立成分は，γ プロテオバクテリアのグループに相当していたことを思い出してほしい．高得点の上位遺伝子産物群を，lipopolysaccharide biosynthesis pathway にマッピングした．図 2.6 で，太線で強調された遺伝子群が，第4独立成分で高得点の遺伝子だったものである．図 2.7 で，灰色で示した遺伝子が，大腸菌 *Escherichia coli* が持っている遺伝子群である．2つの図において，図 2.6 でマークしている遺伝子は，図 2.7 でマークされている遺伝子に含まれており，対応していることが分かる．大腸菌は，γ プロテオバクテリアのグループの生物種であるので，これによって，抽出した軸の生物種グループとしての妥当性が確認できた．またこれらの高得点の遺伝子は，他の生物種グループにおいては，あまり存在しないという傾向を示した．もちろんこれは一例であり，他の特異的パスウェイの抽出も可能であり，他の生物種グループの特異的なパスウェイの抽出も可能である．

また生物種グループを表す各独立成分において，高得点の遺伝子が，どのような機能に対応しているかを検証した．以下に例を示す．fungi のグループを表す第2独立成分において高得点の遺伝子群は，DNA-directed RNApolymerase III subunit とアノテーションされている遺伝子が多かった．古細菌のグループを表す第3独立成分において高得点の遺伝子群は，ribosomal proteins 30S and 50S とアノテーションされている遺伝子が多かった．シアノバクテリアのグループを表す第9独立成分において高得点の遺伝子群は，photosystem とアノテーションされている遺伝子が多く，光合成特異的な遺伝子の獲得パターンを持つグループであった．このように，抽出した生物種グループと機能との関係は，これまでの生物学的に確認されている事実と一致し，手法の妥当性が確認できた．

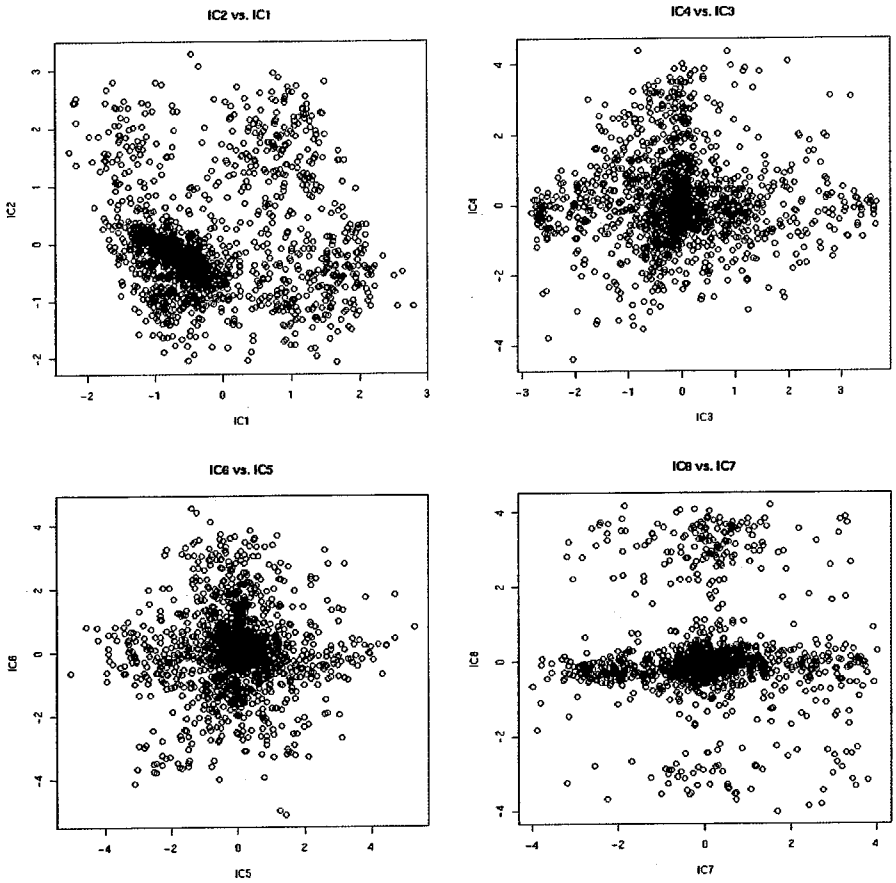

図 2.5: 独立成分スコアの散布図：
第 1 独立成分と第 2 独立成分 (IC1 vs. IC2) の散布図 (左上), 第 3 独立成分と第 4 独立成分 (IC3 vs. IC4) の散布図 (右上)，第 5 独立成分と第 6 独立成分 (IC5 vs. IC6) の散布図 (左下), 第 7 独立成分と第 8 独立成分 (IC7 vs. IC8) の散布図 (右下).

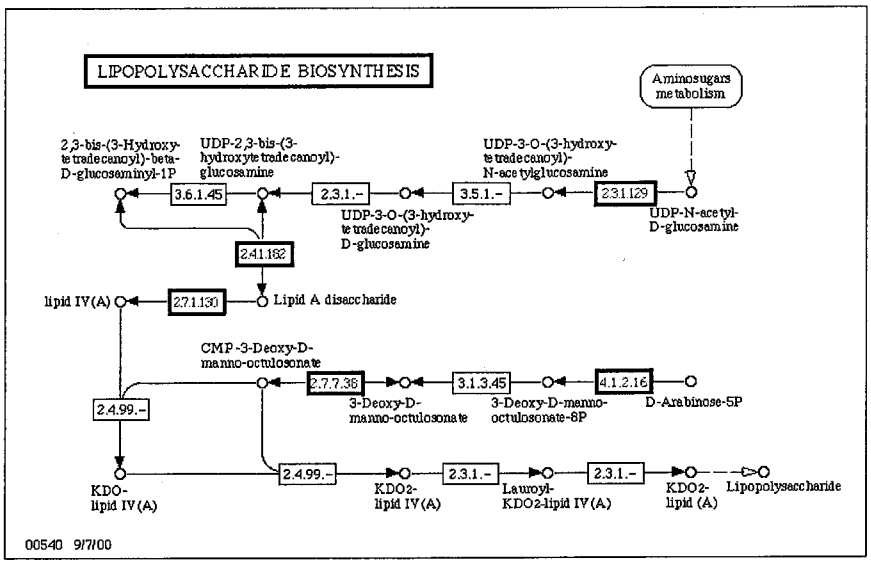

図 2.6: γ プロテオバクテリアのグループ特異的として抽出されたパスウェイの例：独立成分 IC4 で高いスコアを持つ遺伝子群は実線で示されている.

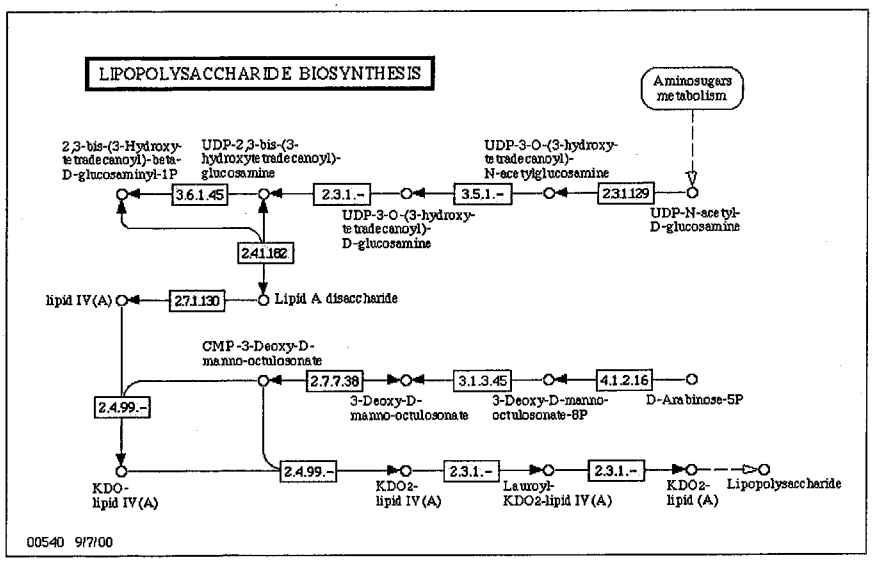

図 2.7: γ プロテオバクテリアに属する大腸菌のパスウェイの例：γ プロテオバクテリアに属する大腸菌の遺伝子群は灰色で示されている.

2.4 考察

本研究では，独立成分分析という統計手法を用いた系統プロファイルの解析法を提案した．2875個のオーソログ遺伝子，77生物種から構成された系統プロファイルの独立成分分析により，真核生物(animal)，真核生物(plant/fungi)，古細菌，プロテオバクテリア(γ)，プロテオバクテリア(δとε)，プロテオバクテリア(α)，グラム陽性菌，クラミジア，シアノバクテリアの9つの生物種グループを表す独立成分を抽出できた．つまり，この77種の生物種の間では，遺伝子の得失のパターンは主に9つのパターンで説明できることを意味する．また，元のデータで，77個の生物種が9個の生物種グループに単純化できていることになる．今後，全ゲノム配列が決まった生物種は，かなりのペースで増えていくであろう．通常の解析では生物種が溢れるとデータの解析が困難になるであろうが，提案手法により，遺伝子の得失のパターンの解析を容易に行え，それに基づく生物種の分類を行うことが期待できる．

提案手法のもう一つの特長は，各生物種グループへの遺伝子の帰属度を，数量化できるという点である．つまり，全生物種で表された多次元の空間から，主要な生物種グループで表された小次元の空間へ数学的に射影することによって，生物種グループとしての視点から遺伝子間の関係を捉えることができる．各生物種グループ軸において，全遺伝子の分布を視覚化することができ，ある生物グループを特徴づける遺伝子群を検出することを可能にした．射影後の空間において高得点を示す遺伝子群は，各生物種グループにおいて，排他的に存在し，低得点の遺伝子群は排他的に存在しないという傾向をみせることが分かった．実際に，特徴的な遺伝子群と生物的機能との対応を調べた結果，一例として，γプロテオバクテリアに特徴的な遺伝子が多く作用している代謝パスウェイを確認できた．

各生物種グループに特異的な遺伝子群と生物学的な機能を確認したところ，例えば，シアノバクテリアグループに帰属度が高いとされた遺伝子は，光合成の機能を持っている遺伝子であったりなど，機能と生物種グループとの対応の妥当性が確認できた．これは，既知の生物学的な事実と，生物種グループごとの遺伝子の有る無しのパターンを結びつけることによって，形質的な生物種グループの特徴は既知であるが，どの遺伝子がその形質に相当するかは同定されていないという遺伝子を同定できる可能性が考えられる．逆に言えば，現在は機能未知の遺伝子群が，それらの既知の生物種グループの性質を表すものであるということが，生物種グループに対する遺伝子の帰属度から推定できる可能性が考えられる．本研究では，その過程まではいけなかったが，今後の検討課題である．

また抽出された独立成分と主要な生物種グループの対応に注目し，独立成分と生物種との相関係数を基にした新しい系統分類法を開発した．クラスター分析を

適用したところ，既存の方法では，古細菌は真正細菌に近いという階層構造を示したが，提案した方法では，古細菌は真核生物に近いという階層構造を示した．また，より簡単で分かりやすい階層構造を示すことができ，分子生物学の分野で主張されている分類基準を支持する結果となった．生物学的に意味があると思われる独立成分の数の決め方という問題点は残されているが，提案手法では，独立成分分析によって生物学的に意味のある成分を基にクラスタリングを行ったことによって，分類法の性能が向上したと解釈することができる．

表 2.5: 各生物種の独立成分スコアが高い遺伝子数のリストI：
独立成分スコアの上位5%に，各生物種の遺伝子が何個占めているかを表したもの
($s_j > 0.05$ 分位点)．

No.	Abbr.	Organism	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9
1	hsa	Homo sapiens	143	23	64	28	29	41	33	33	18
2	dme	Drosophila melanogaster	143	30	63	23	23	27	23	31	18
3	cel	Caenorhabditis elegans	139	34	59	24	28	26	25	30	14
4	ath	Arabidopsis thaliana	1	48	34	9	20	18	18	40	23
5	sce	Saccharomyces cerevisiae	91	38	70	24	34	26	32	53	15
6	spo	Schizosaccharomyces pombe	84	42	66	20	28	21	26	46	16
7	mja	Methanococcus jannaschii	8	5	133	16	36	17	17	49	22
8	mth	Methanobacterium thermoautotrophicum	8	8	131	19	34	18	16	52	18
9	afu	Archaeoglobus fulgidus	11	6	136	23	54	29	26	59	26
10	mka	Methanopyrus kandleri	13	6	119	27	45	32	32	50	15
11	tac	Thermoplasma acidophilum	15	1	117	6	42	13	27	49	15
12	tvo	Thermoplasma volcanium	14	2	117	6	35	13	26	47	15
13	pho	Pyrococcus horikoshii	7	2	129	8	41	22	18	30	14
14	pab	Pyrococcus abyssi	7	6	137	14	45	23	23	39	16
15	pfu	Pyrococcus furiosus	9	4	138	8	40	22	22	45	18
16	ape	Aeropyrum pernix	10	5	119	13	38	22	26	51	16
17	sso	Sulfolobus solfataricus	16	2	132	5	23	15	26	51	23
18	sto	Sulfolobus tokodaii	13	4	131	4	27	16	25	53	20
19	pai	Pyrobaculum aerophilum	14	0	130	11	37	19	32	56	22
20	eco	Escherichia coli K-12 MG1655	33	24	31	142	106	86	93	121	48
21	ecj	Escherichia coli K-12 W3110	33	23	31	142	96	86	91	121	47
22	ece	Escherichia coli O157 EDL933	33	26	31	141	102	85	94	130	49
23	ecs	Escherichia coli O157 Sakai	33	26	31	141	100	84	92	129	46
24	sty	Salmonella typhi	32	24	33	140	95	76	95	127	48
25	stm	Salmonella typhimurium	33	24	34	141	94	77	93	128	48
26	ype	Yersinia pestis	31	23	27	136	88	74	91	128	39
27	hin	Haemophilus influenzae	32	14	13	127	75	54	67	105	23
28	pmu	Pasteurella multocida	32	11	15	138	72	45	74	114	29
29	xfa	Xylella fastidiosa	22	11	14	62	29	30	39	98	25
30	vch	Vibrio cholerae	36	16	31	132	79	48	88	117	37
31	pae	Pseudomonas aeruginosa	34	19	26	120	87	53	71	131	43
32	buc	Buchnera sp. APS	13	8	10	18	11	6	26	56	7
33	nme	Neisseria meningitidis MC58	20	14	14	84	50	10	50	109	27
34	nma	Neisseria meningitidis Z2491	21	14	14	85	49	12	50	109	29
35	rso	Ralstonia solanacearum	28	24	28	83	82	23	64	121	34
36	hpy	Helicobacter pylori 26695	21	15	15	53	126	16	29	98	24
37	hpj	Helicobacter pylori J99	21	15	14	53	123	16	27	98	24
38	cje	Campylobacter jejuni	20	9	19	61	89	18	28	105	29
39	rpr	Rickettsia prowazekii	15	3	6	32	10	36	13	76	13
40	rco	Rickettsia conorii	18	2	6	32	11	38	13	76	14

表 2.6: 各生物種の独立成分スコアが高い遺伝子数のリストII：
独立成分スコアの上位5%に，各生物種の遺伝子が何個占めているかを表したもの
($s_j > 0.05$ 分位点).

No.	Abbr.	Organism	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9
41	mlo	Mesorhizobium loti	33	21	25	84	80	123	64	125	39
42	sme	Sinorhizobium meliloti	30	20	23	98	87	138	68	116	45
43	atu	Agrobacterium tumefaciens C58 (UWash)	29	21	20	94	84	140	65	118	37
44	atc	Agrobacterium tumefaciens C58 (Cereon)	29	19	20	94	82	141	65	118	37
45	bme	Brucella melitensis	29	28	23	85	77	100	63	112	34
46	ccr	Caulobacter crescentus	25	13	15	68	56	60	51	105	22
47	bsu	Bacillus subtilis	34	18	19	17	82	55	102	108	24
48	bha	Bacillus halodurans	33	14	21	14	76	56	94	107	30
49	sau	Staphylococcus aureus N315 (MRSA)	26	17	16	14	83	31	137	89	7
50	sav	Staphylococcus aureus Mu50 (VRSA)	26	17	16	14	81	31	136	87	7
51	lmo	Listeria monocytogenes	28	13	17	8	67	30	115	82	17
52	lin	Listeria innocua	28	13	17	9	68	30	110	81	16
53	lla	Lactococcus lactis	27	12	12	12	61	34	91	72	12
54	spy	Streptococcus pyogenes SF370	24	10	15	15	59	19	100	62	6
55	spn	Streptococcus pyogenes MGAS8232	21	10	16	16	60	29	104	81	15
56	spr	Streptococcus pneumoniae R6	21	10	8	17	61	32	103	74	15
57	cac	Clostridium acetobutylicum	23	11	20	17	82	38	81	84	24
58	cpe	Clostridium perfringens	29	6	25	22	71	35	81	86	20
59	mge	Mycoplasma genitalium	8	4	4	9	15	9	25	0	4
60	mpn	Mycoplasma pneumoniae	8	4	3	11	19	9	31	0	4
61	mpu	Mycoplasma pulmonis	11	4	3	12	20	15	31	0	6
62	uur	Ureaplasma urealyticum	7	7	3	11	21	7	25	0	6
63	mtu	Mycobacterium tuberculosis H37Rv	28	17	22	23	85	30	5	93	0
64	mtc	Mycobacterium tuberculosis CDC1551	27	16	22	22	85	28	4	92	0
65	mle	Mycobacterium leprae	23	5	9	16	46	19	0	81	0
66	ctr	Chlamydia trachomatis	14	6	11	18	18	13	18	143	13
67	cmu	Chlamydia muridarum	15	6	11	18	17	12	17	141	13
68	cpn	Chlamydophila pneumoniae CWL029	16	6	11	18	23	14	20	143	14
69	cpa	Chlamydophila pneumoniae AR39	15	6	11	18	20	14	20	143	14
70	cpj	Chlamydophila pneumoniae J138	16	6	11	18	22	14	20	143	14
71	bbu	Borrelia burgdorferi	16	7	12	15	28	7	32	30	6
72	tpa	Treponema pallidum	17	9	13	23	28	13	17	39	9
73	syn	Synechocystis sp. PCC6803	30	24	21	46	75	32	35	102	142
74	ana	Anabaena sp. PCC7120	28	21	24	47	75	40	43	100	143
75	dra	Deinococcus radiodurans	28	11	26	31	81	34	43	96	24
76	aae	Aquifex aeolicus	24	10	27	30	55	22	23	101	30
77	tma	Thermotoga maritima	25	9	23	14	49	35	44	68	24

第3章　異質なゲノムデータを用いたオペロンの解析

3.1　序論

原核生物は，パスウェイ上で遺伝子産物が連続的に機能し，ゲノム上で遺伝子が隣接し，遺伝子発現パターンが似ている”オペロン”と呼ばれる特徴的な遺伝子クラスター群を持つことが知られている．ほとんどの原核生物では，オペロンである遺伝子クラスターは，同じ上流プロモータのもとで制御されており，原核生物の転写単位となっている．それゆえ，その構造を理解することは，遺伝子ネットワークやタンパク質間相互作用の仕組みを解明するのに重要である．

例えば，図 3.1 は，よく研究されている大腸菌のトリプトファンオペロンであり，5つの遺伝子がトリプトファンの生合成を行う5つの酵素に翻訳される様子を示す．5個の遺伝子は単一の mRNA 分子として転写され，このとき遺伝子発現は協調して行われる．このような遺伝子クラスターはオペロンと呼ばれる．実験的なオペロンの同定は，コストや時間がかかり，網羅的な同定作業をウェットな実験作業でやり抜くことは非常に難しい [58]．それゆえ，計算機でのオペロン予測が注目を浴びており，配列解析に基づく方法 [61, 44, 16] や，複数のデータのグラフ比較に基づく方法 [38, 66] などが提案されている．

オペロンは，いくつかの生物的属性が凝縮した相関の一つの形であると考えられる．オペロンに属する遺伝子は，染色体上で隣接し，同じような遺伝子発現パターンを示し，その酵素はパスウェイ上で連続的に化学反応を触媒する傾向があることが言われている．逆に考えれば，ゲノム上での遺伝子の並び，遺伝子発現プロファイル，パスウェイ上での化学反応の連続性を表す3つのデータがあれば，これらのデータ間の相関を検証し，オペロンに属する遺伝子を抽出できることが期待できる．

本研究では，カーネル正準相関分析のモデルを拡張することによって，複数の異なるゲノムデータセット間の相関に寄与する遺伝子クラスターを抽出する手法を開発した [63]．実際に，提案した手法を大腸菌のオペロン構造の検出法として適用した．ここでは，代謝および制御パスウェイ上の機能的な遺伝子間の関係，ゲノム上（染色体上）での遺伝子の隣接関係，マイクロアレイ実験で共発現する遺

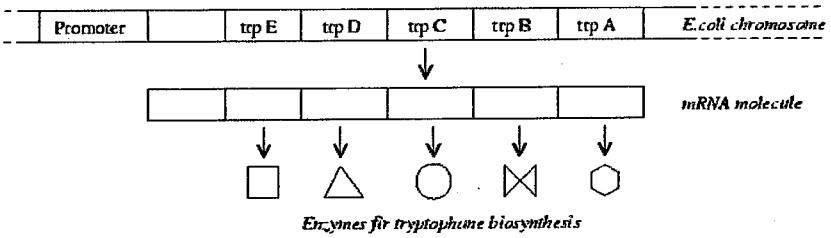

図 3.1: アミノ酸トリプトファン生合成で働く酵素をコードする大腸菌のオペロンの例

伝子間の関係を表す，パスウェイ，ゲノム，発現データの３つのデータを用いた．この研究で用いたデータセットは，KEGG データベース [29, 69] から得た．検出される相関（この場合オペロン）に寄与していると考えられる遺伝子群を抽出した．大腸菌のオペロンデータベースと比較して予測精度を検証した結果，抽出した遺伝子群は，既知のオペロンに属する遺伝子群に対応していることが確認でき，提案する相関解析法の有効性を示した．

3.2 方法

3.2.1 データセット

オペロンに属する遺伝子群は，染色体上で隣接しており，共発現し，代謝パスウェイで連続して働く傾向があることが言われている．そこで，パスウェイ，ゲノム上の並び，発現データの３つのデータセットを用いた．以下では，それぞれ，pathway，genome，expression と単純にラベル付けをして記すことにする．それぞれのデータに関して，共通の 740 個の遺伝子セットを用いた．予測したオペロンと，実際のオペロンの対応を検証するために，正解データとして，大腸菌 *Escherichia coli K-12* のオペロンデータベース [25, 67] を用いた．

パスウェイ (pathway) パスウェイデータは，KEGG/PATHWAY データベースから大腸菌の代謝パスウェイをダウンロードした．そして，大腸菌 *E. coli K-12* の遺伝子を頂点 (ノード)，その遺伝子産物の酵素がパスウェイ上で連続的な化学反応を触媒するとき辺 (エッジ) を結んだグラフを構築した．ここでいうグラフとは，頂点 (ノード) が遺伝子，辺 (エッジ) が遺伝子間の二項関係を表すものである．

ゲノム上の並び (genome) 遺伝子のゲノム上での位置情報は KEGG/GENES データベースからダウンロードした．染色体上での遺伝子の位置情報から，ノー

ドが遺伝子，エッジが染色体上での隣接関係である線形グラフを構築した．

遺伝子発現データ (expression) 遺伝子発現データは，KEGG/EXPRESSION データベース [69] に保存されている，大腸菌 *E. coli K-12* の 48 種類の実験に基づくマイクロアレイのデータを用いた．各アレイで，全遺伝子に対して，赤と緑の蛍光色素でラベルしたプローブから蛍光強度ペアのデータが得られる．ここで，赤色は Cy5 に，緑色は Cy3 にそれぞれ相当し，Cy5 と Cy3 の蛍光強度から遺伝子の発現をモニターする．一般的なマイクロアレイ解析方法 [65] に従い，それぞれの遺伝子に対して，対数比 $\log(R_S - R_B)/(G_S - G_B)$ で発現レベルを測った．ここで，G_B は，バックグラウンドの値，G_S はシグナルの値，R_B はターゲットのバックグラウンドの値，R_S はターゲットのシグナルを表す．各遺伝子に対して，48 種類の数値ベクトルを持つデータとなる．

3.2.2 カーネル法

カーネルの意味 ここで，全てのデータをカーネルで，統一的に表すことを考える．本研究では，オブジェクトは遺伝子に対応するので，カーネルとは，ゲノムデータが与えられたときの，遺伝子ペアペアの類似度の尺度であると解釈することができる．つまり，遺伝子 $\mathbf{x}$ と遺伝子 $\mathbf{x}'$ が与えられたとき，そのカーネル $k(\mathbf{x}, \mathbf{x}')$ は，遺伝子 $\mathbf{x}$ と遺伝子 $\mathbf{x}'$ の類似度になる．

数値ベクトルのカーネル データセットが expression のように数値データの場合，例えば，ガウシアンカーネルなどを用いて，以下のようにカーネル行列を計算することができる．

$$\begin{aligned}(K)_{ij} &:= k_{\text{gaussian}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \\ &= \exp(- \parallel \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \parallel^2 / \sigma^2).\end{aligned}$$

グラフのカーネル データセットが pathway のようにグラフ構造の場合，拡散カーネル [30] というカーネル関数を使うことによって，グラフをカーネル行列に変換できる．重み無しの無向グラフ $\Gamma = (V, E)$ が与えられたとすると，そのグラフのラプラシアン行列 L は，以下のように表せる．

$$L_{ij} = \begin{cases} -1 & \text{for } i \sim j, \\ d_i & \text{for } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

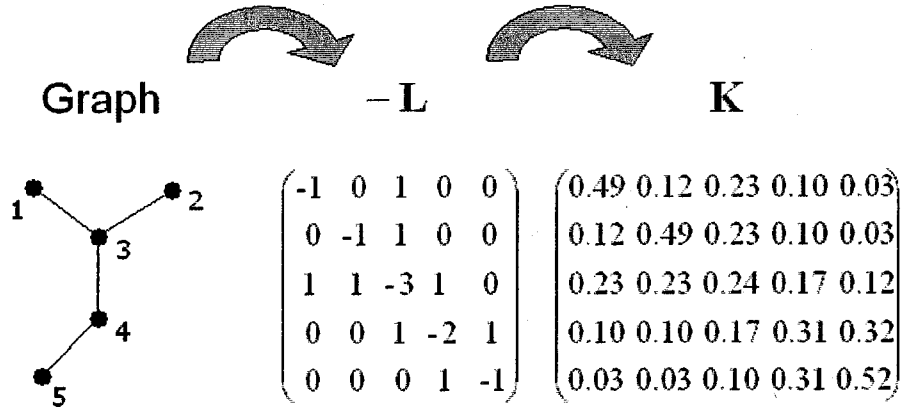

図 3.2: 拡散カーネルの例：グラフ上のノード間の類似度を計算

.

ここで，$i \sim j$ は，i番目と j番目の遺伝子がグラフ上 (例えばパスウェイ上) において隣接関係にあることを示し，d_i は i番目の対象が持つエッジ数 (隣接している遺伝子数) を示す．このとき，その L の行列指数演算

$$K = \exp(-\beta L) = \lim_{m \to \infty} \left(I + \frac{\beta L}{m} \right)^m$$

は，対称行列かつ正定値行列となり（β は正の定数），カーネル行列としての性質を満たすことが知られている．ここで，I は，対角成分が1で，それ以外はゼロの要素を持つ単位行列を表す．図 3.2 は，グラフのデータから，類似度行列であるカーネル行列と，変換されるまでの一例を示している．

ゲノムデータのカーネル化　パスウェイデータのデータ構造はグラフ，ゲノム上での遺伝子の隣接関係は線形グラフと見なせるので，これらのデータは，拡散カーネルを使ってデータをカーネルに変換する．遺伝子発現データの発現プロファイルは，一つの遺伝子に対して数値ベクトルとして得られるので，ガウシアンカーネルを用いて，カーネルに変換する．その結果，740×740 の大きさのカーネル行列を3つのデータに対して得ることができ，それぞれ $K_{pathway}$，K_{genome}，$K_{expression}$ と表す．

3.2.3　カーネル正準相関分析

同じ対象 (ここでは遺伝子に相当する) に関して2種類の異質なデータセットが与えられたとき，そのデータセット間の相関関係を解析したいとする．両方のデー

タセットが数値ベクトルで構成されている場合，多変量解析法の一つである正準相関分析 (canonical correlation analysis (CCA)) [21] が有効な手法として知られている．両データセット間の相関を最もよく表すような新しい特徴量に変換し，それによって2つのデータ間の相関について考えようとするのが，通常の正準相関分析の考え方である．しかしながら，データが数値データではなく，グラフ構造，文字列といった離散データや構造データの場合には，直接的には適用することはできない．そこで，データ構造の異なるデータセット間の相関解析の枠組みが必要になる．本研究では，この問題に対し，カーネル法のアイデアを用いて，通常の正準相関分析のモデルを一般化させたカーネル正準相関分析 (kernel canonical correlation analysis (KCCA))[1, 6] を使った相関解析法の開発を試みる．

通常のカーネル正準相関分析 (OKCCA) この節では，通常のカーネル正準相関分析 (ordinary kernel canonical correlation analysis (OKCCA)) について簡単に説明する．詳細は，参考文献 [1, 6] を参照されたい．この方法の目的は，2種類のデータセット $\{\mathbf{x}_1^{(i)}\}_{i=1}^n$，$\{\mathbf{x}_2^{(i)}\}_{i=1}^n$ の間の相関を検出することにある．ここで，n は，遺伝子の数，データ $\mathbf{x}_1^{(i)}$ と $\mathbf{x}_2^{(i)}$ は，集合 $\mathcal{X}_1$ と $\mathcal{X}_2$ に属する．ここでは，各データは遺伝子の情報の一つの表現形式であると考えることができる．例えば，もし $\mathcal{X}_1$ が塩基配列の集合で，$\mathcal{X}_2$ が遺伝子発現プロファイルの集合のとき，$\mathbf{x}_1^{(i)}$ は i 番目の遺伝子の塩基配列になるであろうし，$\mathbf{x}_2^{(i)}$ は，i 番目の遺伝子発現プロファイルになる．

同じ遺伝子に関する2種類のデータセット $\{\mathbf{x}_1^{(i)}\}_{i=1}^n$，$\{\mathbf{x}_2^{(i)}\}_{i=1}^n$ が与えられたとする．遺伝子 $\mathbf{x}_1^{(i)}$，$\mathbf{x}_2^{(i)}$ が，あるヒルベルト空間 H_1，H_2 に，$\phi_1 : \mathcal{X}_1 \to H_1$，$\phi_2 : \mathcal{X}_2 \to H_2$ によってそれぞれ写像されたとき，通常の正準相関分析の概念を，写像後のセット $\{\phi_1(\mathbf{x}_1^{(i)})\}_{i=1}^n$，$\{\phi_2(\mathbf{x}_2^{(i)})\}_{i=1}^n$ に適用することを考える．各 $\mathbf{x}_1^{(i)}$，$\mathbf{x}_2^{(i)}$ に対して，次のような特徴量をそれぞれ定義する．

$$\begin{aligned} u_1^{(i)} &:= \langle f_1, \phi_1(\mathbf{x}_1^{(i)})\rangle, \\ u_2^{(i)} &:= \langle f_2, \phi_2(\mathbf{x}_2^{(i)})\rangle. \end{aligned}$$

ここで，$\langle\cdot,\cdot\rangle$ は，ヒルベルト空間における内積を表す．また，その標本平均，標本分散，標本共分散を，それぞれ

$$\begin{aligned} \bar{u}_k &:= \frac{1}{n}\sum_{i=1}^n u_k^{(i)} \quad (k=1,2), \\ var(u_k) &:= \frac{1}{n}\sum_{i=1}^n (u_k^{(i)} - \bar{u}_k)^2 \quad (k=1,2), \\ cov(u_1,u_2) &:= \frac{1}{n}\sum_{i=1}^n (u_1^{(i)} - \bar{u}_1)(u_2^{(i)} - \bar{u}_2), \end{aligned}$$

とおく．これら2つの特徴量 u_1，u_2 の標本相関係数

$$corr(u_1, u_2) := \frac{cov(u_1, u_2)}{(var(u_1)var(u_2))^{1/2}}$$

を最大にするような，2つの軸 $f_1 \in H_1$ と $f_2 \in H_2$ を求めることが，カーネル正準相関分析の目的となる．しかしながら，H_1 と H_2 の次元はサンプル数よりも大きく，この解は一意には決まらないので，正則化の概念を導入する必要がある．一つの正則化の方法としては，ヒルベルト空間における f_1 と f_2 のノルムをペナルティとして付加し，以下のようなペナルティ付きの最大化問題を考える．

$$\frac{cov(u_1, u_2)}{(var(u_1) + \lambda_1 \parallel f_1 \parallel^2)^{\frac{1}{2}}(var(u_2) + \lambda_2 \parallel f_2 \parallel^2)^{\frac{1}{2}}}.$$

ここで，λ_1，λ_2 は正則化パラメータを表し，標本相関係数の最大化と，ノルム $\parallel f_k \parallel$ $(k = 1, 2)$ を小さく抑えることを制御する．また，$\mathbf{x}_1$，$\mathbf{x}_2$ のカーネル行列として，

$$\begin{aligned}(K_1)_{ij} &:= k_1(\mathbf{x}_1^{(i)}, \mathbf{x}_1^{(j)}) = \langle \phi_1(\mathbf{x}_1^{(i)}), \phi_1(\mathbf{x}_1^{(j)}) \rangle \\ (K_2)_{ij} &:= k_2(\mathbf{x}_2^{(i)}, \mathbf{x}_2^{(j)}) = \langle \phi_2(\mathbf{x}_2^{(i)}), \phi_2(\mathbf{x}_2^{(j)}) \rangle\end{aligned}$$

が，$i, j = 1, 2, \cdots, n$ に対して得られているとき，上のペナルティ付きの最大化問題は，最終的に以下の一般化固有値問題に帰着する（証明の詳細は参考文献 [47] を参照）．

$$\begin{pmatrix} \mathbf{0} & K_1K_2 \\ K_2K_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} = \rho \begin{pmatrix} (K_1 + \frac{n\lambda_1}{2}I)^2 & \mathbf{0} \\ \mathbf{0} & (K_2 + \frac{n\lambda_2}{2}I)^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}.$$

ここで，I は単位行列を表す．この固有ベクトル $\boldsymbol{\alpha}_1 = (\alpha_{11}, \cdots, \alpha_{1n})^\top$，$\boldsymbol{\alpha}_2 = (\alpha_{21}, \cdots, \alpha_{2n})^\top$ を用いて，正準相関得点 (canonical correlation score) は，

$$\begin{aligned}u_1^{(i)} &= \sum_{j=1}^{n} K_{1ij}\alpha_{1j}, \\ u_2^{(i)} &= \sum_{j=1}^{n} K_{2ij}\alpha_{2j},\end{aligned}$$

と求めることができる．つまり，実際には，$\phi_1(\mathbf{x}_1^{(i)})$，$\phi_2(\mathbf{x}_2^{(i)})$ を計算をすることなく，カーネル行列を入力として正準相関得点を計算できる点が特長である．これにより，対象間の類似度行列を表現するカーネル行列があれば，数値データ以外の離散データにも適用可能であるので，どんなカーネル関数を用いてデータのカーネル行列を得るかということが本質的な問題となる．

データセットが3つ以上の場合のカーネル正準相関分析 (MKCCA) 本研究では，データセットの種類が三つ以上与えられた場合も，同様の解析が可能になるようにモデルの拡張を行った．ここでは，P個の種類のデータセットがあるとする．これを，ここでは，多重カーネル正準相関分析 (multiple kernel canonical correlation analysis (MKCCA)) と呼ぶことにする．

同じ遺伝子に関するP種類のデータセット $\{\mathbf{x}_p^{(i)}\}_{i=1}^n$ $(p = 1, 2, ..., P)$ が与えられたとする．また，$\mathbf{x}_p$ のカーネル行列として，

$$(K_p)_{ij} \;:=\; k_p(\mathbf{x}_p^{(i)}, \mathbf{x}_p^{(j)})$$

が，$i, j = 1, 2, \cdots, n$ に対して得られているとき，P個のデータセット間の相関最大化問題は，最終的に以下の一般化固有値問題に帰着する（証明の詳細は参考文献 [47] を参照）．

$$\begin{pmatrix} \mathbf{0} & K_1K_2 & \dots & K_1K_2 \\ K_2K_1 & \mathbf{0} & \dots & K_1K_2 \\ \vdots & \vdots & \ddots & \vdots \\ K_2K_1 & \dots & K_2K_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_P \end{pmatrix} =$$

$$\rho \begin{pmatrix} (K_1 + \frac{n\lambda_1}{2} I)^2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & (K_1 + \frac{n\lambda_1}{2} I)^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (K_2 + \frac{n\lambda_2}{2} I)^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_P \end{pmatrix}.$$

ここで，I は単位行列を表す．この固有ベクトル $\boldsymbol{\alpha}_p = (\alpha_{p1}, \cdots, \alpha_{pn})^\mathsf{T}$ を用いて，正準相関得点は，

$$u_p^{(i)} \;=\; \sum_{j=1}^{n} K_{pij}\alpha_{pj}$$

と求めることができる．

データ融合を組み合わせたカーネル正準相関分析 (IKCCA) 複数のデータの同時相関を最大にすることは，制約が強すぎて，生物的に意味がある特徴を抽出できない場合がある．ここでは，異質なデータの融合を取り入れたカーネル正準相関分析 (integrated kernel canonical correlation analysis (IKCCA)) を提案した．カーネル正準相関分析の文脈におけるデータ融合は，本研究が，初めての報告となる．

ここで，$P \geq 1$ 個の異質なゲノムデータが得られており，それぞれ P 個のカーネル $K_1, \ldots, K_P$ で表されているとする．K_p は p 番目のデータセットに関する，

遺伝子間の類似度行列を表す．一つの統合法として，$K^* = \sum_{p=1}^{P} K_p$ と和を取ることでデータの統合をすることを提案する．簡単ではあるが，この方法の実際の有効性は確かめられている [40, 63]．それぞれのカーネル行列は，各ゲノムデータに基づく遺伝子間の類似度行列を表すので，これらの和を取ることは，多くのゲノムデータで高い類似度を示す遺伝子ペアほど，ペア間の強さがより強調される．少ないゲノムデータでしか高い類似度を示さない遺伝子ペアは，ペア間の強さは小さく抑えられ，ノイズを抑える効果が期待でき，より信頼性のある遺伝子間の類似度行列を構築することが期待できる．

カーネル正準相関分析への応用のため，より定式的にこの問題を考える．2つのメインの属性 x と y があり，x はいくつかの部分クラス $x_p\,(p = 1, 2, \cdots, P)$ を持ち，y もいくつかの部分クラス $y_q\,(q = 1, 2, \cdots, Q)$ を持っているとする．部分クラス毎にカーネル行列 $\{K_{x_1}, \cdots, K_{x_P}\}$，$\left\{K_{y_1}, \cdots, K_{y_Q}\right\}$ が得られていると仮定し，メインのクラス x と y に対する統合したカーネル行列を，次のように定義する．

$$K_{\mathcal{P}}^* = \sum_{p=1}^{P} K_{x_p}, \quad K_{\mathcal{Q}}^* = \sum_{q=1}^{Q} K_{y_q}.$$

ここで，統合したカーネル行列を入力として，通常のカーネル正準相関分析の枠組みで解析しようというのが，ここでのアイデアであり，以下の一般化固有値問題に置き換えることができる．

$$\begin{pmatrix} \mathbf{0} & \sum_p K_{x_p} \sum_q K_{y_q} \\ \sum_q K_{y_q} \sum_p K_{x_p} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_y \end{pmatrix}$$

$$= \rho \begin{pmatrix} (\sum_p K_{x_p} + \lambda_x I_x)^2 & \mathbf{0} \\ \mathbf{0} & (\sum_q K_{y_q} + \lambda_y I_y)^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_y \end{pmatrix} .$$

同様に，正準相関得点 (canonical correlation score) は，

$$\begin{aligned} u_{\mathcal{P}}^{(i)} &= \sum_{j=1}^{n} K_{\mathcal{P}ij}^* \alpha_{1j}, \\ u_{\mathcal{Q}}^{(i)} &= \sum_{j=1}^{n} K_{\mathcal{Q}ij}^* \alpha_{2j}, \end{aligned}$$

と求めることができる．

3.2.4 遺伝子相関クラスターの抽出

カーネル正準相関分析の出力である正準相関得点を使って，データ間の相関に寄与する遺伝子群を抽出する手順を提案する．2つのデータを比較したとき，2種

類の正準変量があり，3つのデータを比較したとき，3種類の正準変量がある．ここで，複数の種類の正準変量間を平均化および絶対値を取ることにより，相関への寄与を計るスコア $s \in R^n$ を，$i = 1, \cdots, n$ に対して，次のように定義する．

$$s_{OKCCA}(i) = \left| \frac{u_1^{(i)} + u_2^{(i)}}{2} \right|, \quad s_{MKCCA}(i) = \left| \frac{1}{P} \sum_{j=1}^{P} u_j^{(i)} \right|, \quad s_{IKCCA}(i) = \left| \frac{u_{\mathcal{P}}^{(i)} + u_{\mathcal{Q}}^{(i)}}{2} \right|.$$

ここで，OKCCA はデータセットが2つの場合の通常のカーネル正準相関分析，MKCCA はデータセットが3つ以上の場合のカーネル正準相関分析，IKCCA はデータ融合を伴うカーネル正準相関分析を意味している．

このスコアの解釈であるが，例えば，$s(i) = 0$ は，i 番目の遺伝子が根底にあるデータ間の相関にほとんど寄与していないことを表し，逆に $s(i)$ のスコアが高ければ，i 番目の遺伝子がデータ間の相関に寄与していることを表す．正準相関スコアが何らかの生物学的な特徴に起因している場合，高い正準スコアを持つ遺伝子群は，低い遺伝子群よりも，より相関に寄与していることを表す．ここでは，遺伝子群の検出が目的であるので，このスコアに対してある閾値を設定し，それ以上のスコアを取る遺伝子群を選んでやるという方針をとった．

3.3 結果

3.3.1 オペロンに属する遺伝子の検出

実際に，提案した手法を大腸菌のオペロン構造の検出法として適用した．オペロン検出に最適な生物学的属性の組合せを調べるため，様々なデータの組合せの比較を行った．

まず，通常のカーネル正準相関分析 (OKCCA) を全てのペアワイズなデータの組合せに対して適用した．次に，多重カーネル正準相関分析 (MKCCA) を3つのデータセットに同時に適用した．最後に，データ融合を組み合わせたカーネル正準相関分析 (IKCCA) を可能な全てのデータの組合せに対して適用した．また，データ比較やデータ融合の効果を確認するため，単一のデータに対して，カーネル主成分分析 [45] を適用し，同等の手順で遺伝子抽出を行った．全ての適用手順を，表 3.1 にまとめた．

3.3.2 データの組み合わせによる性能の比較

それぞれの正準相関分析を適用後，最も強い相関である第1正準相関に注目し，これがオペロンに対応していると考えた．図 3.3 は，3つのデータを同時に解析し

表 3.1: オペロン検出のために行った全ての実行例リスト：
通常のカーネル正準相関分析 (OKCCA)，データ融合を伴うカーネル正準相関分析 (IKCCA) は，2つの"Kernel 1" と"Kernel 2" のカーネル行列を入力として実行する．多重カーネル正準相関分析 (MKCCA) は，3つのカーネル行列を入力として実行する．カーネル主成分分析は，"Kernel 1" で表された単一のデータのカーネル行列だけを入力として実行する．

表記法	省略名	手法	Kernel 1	Kernel 2	Kernel 3
OKCCA-a	O-a	KCCA	$K_{pathway}$	K_{genome}	-
OKCCA-b	O-b	KCCA	K_{genome}	$K_{expression}$	-
OKCCA-c	O-c	KCCA	$K_{expression}$	$K_{pathway}$	-
MKCCA	M	MKCCA	$K_{pathway}$	K_{genome}	$K_{expression}$
IKCCA-a	I-a	IKCCA	$K_{genome} + K_{expression}$	$K_{pathway}$	-
IKCCA-b	I-b	IKCCA	$K_{expression} + K_{pathway}$	K_{genome}	-
IKCCA-c	I-c	IKCCA	$K_{pathway} + K_{genome}$	$K_{epression}$	-
KPCA-a	S-a	KPCA	$K_{pathway}$	-	-
KPCA-b	S-b	KPCA	K_{genome}	-	-
KPCA-c	S-c	KPCA	$K_{expression}$	-	-

た多重正準相関分析 (MKCCA) の，第一正準相関の生スコアのクロス散布図を示したものである．図 3.4 は，データ融合をした場合のカーネル正準相関分析を適用したときの，正準相関得点の散布図を IKCCA-a,IKCCA-b,IKCCA-c に対してそれぞれ適用したときのものである．これらの図の中で，丸は一つの遺伝子に対応する．ほとんどの遺伝子は，原点からまっすぐな 45 度の直線上に分布していることが分かる．つまり，異なる生物学的属性間において何らかの相関が検出されていることを意味する．この検出された相関は，高得点または低得点の遺伝子に起因していることから，3.2.4 節で定義した相関遺伝子抽出のためのスコアに，絶対値を取って変換した．

オペロンに属する遺伝子群は，考えている3つの生物学的属性に関して同時に，相関クラスターを構成すると考えられる．ここでは，第一正準相関に寄与している遺伝子群が，オペロンを構成する遺伝子群であると考えられ，他の遺伝子よりも正準相関に強く寄与していると考えられる．これが，3.2.4 節で説明した上位スコアの遺伝子を選択する動機でもあり，オペロンに属する遺伝子群を検出するのに有効であると考えられる．ここでは，異なるオペロン毎に境界を付けるのではなく，オペロンに属する遺伝子群を，オペロンに属さない遺伝子群から分離することを目的としている．

次に，3.2.4 節で説明した手順にしたがい，得られた正準得点や主成分得点の絶

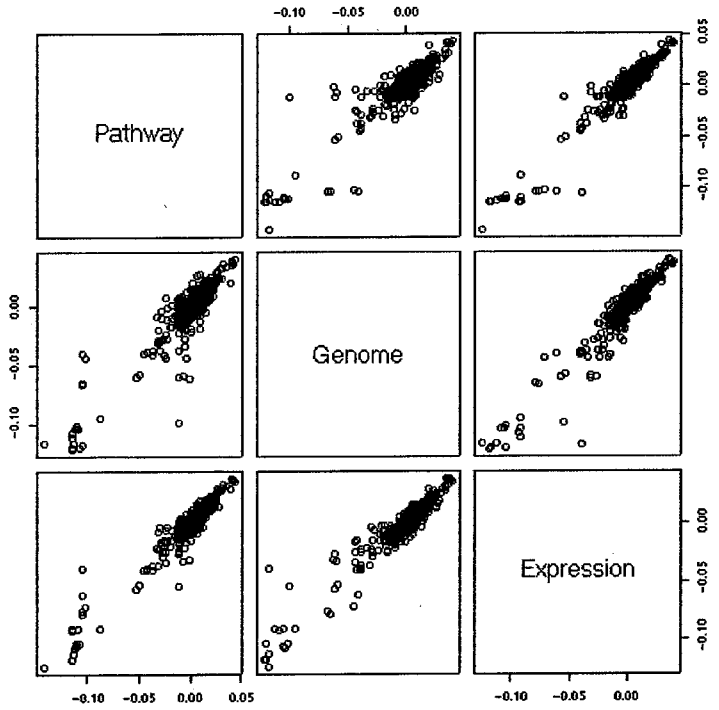

図 3.3: MKCCA の第一正準得点のスコアの多重散布図：
パスウェイ (pathway)，ゲノム上での並び (genome)，発現データ (expression) に対して MKCCA を適用した結果．2行目1列目のパネルは，pathway と genome の第1正準得点の散布図，3行目1列目のパネルは，pathway と expression の第1正準得点の散布図，3行目2列目のパネルは，expression と genome の第1正準得点の散布図を表す．

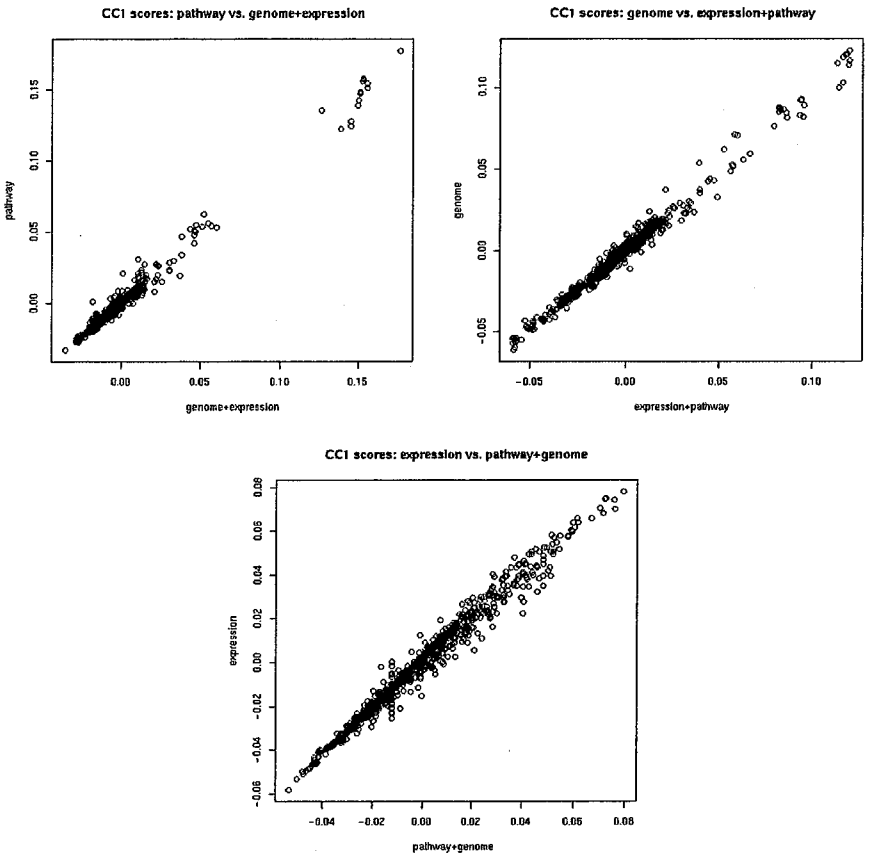

図 3.4: IKCCA の第一正準得点のスコアの散布図：
IKCCA-a は， pathway vs. genome + expression の散布図 (左上)，IKCCA-b は，genome vs. expression + pathway の散布図 (右上)，IKCCA-c は， expression vs. pathway + genome の散布図 (下) を表す.

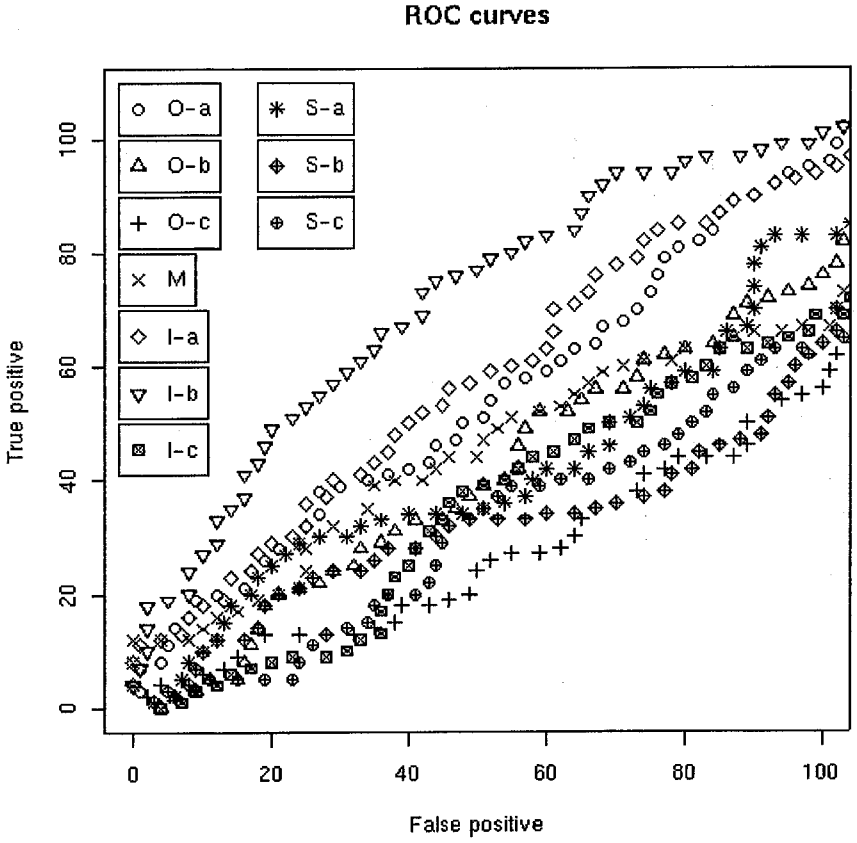

図 3.5: 各 KCCA におけるオペロン遺伝子の検出精度を表す ROC カーブ：正準得点の分位点の閾値を少しずつ変化させて，ROC カーブを描画した．"O-a" は，OKCCA-a, "O-b" は，OKCCA-b, "O-c" は，OKCCA-c, "M" は，MKCCA, "I-a" は，IKCCA-a, "I-b" は，IKCCA-a, "I-c" は，IKCCA-a, "S-a" は，KPCA-a, "S-b" は，KPCA-b, "S-c" は，KPCA-c をそれぞれ表す．X 軸は，false positives の個数を表し，Y 軸は，true positives の個数を表す．

表 3.2: オペロン遺伝子として正確に検出された遺伝子数のリストの一部：ここでは，分位点の閾値を上位の高得点遺伝子の 10% とした (つまり，740 遺伝子中の 74 遺伝子を抽出したときの結果) ときの結果を示す.

オペロン (属する遺伝子数)	O-a	O-b	O-c	M	I-a	I-b	I-c
Biotin metabolism (3)	3	1	0	3	3	3	0
Fatty acid (short-chain) metabolisms (3)	0	3	0	2	0	3	3
Fumarate reductase (4)	4	0	2	4	4	4	0
Galactose metabolism (4)	4	0	0	4	3	4	1
Glycerol-3-phosphate dehydrogenase (3)	0	3	3	3	3	3	3
Menaquinone (vitamin K2) biosynthesis (5)	0	3	0	0	4	0	0
NADH dehydrogenase (13)	0	0	0	0	0	13	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total number (280)	39	34	27	37	42	52	28

対値のスコアが高い遺伝子群を，閾値を少しずつ変化させながら選んでいった．閾値ごとに，この方法で選んだ遺伝子群と，既知のオペロン遺伝子のデータベース [25, 67] と比較し，実際に既知のオペロン遺伝子に対応していた数 (true positives の数) と，既知のオペロン遺伝子に対応していなかった数 (false positives の数) を記録していった．変化させていった閾値の値に基づき，false positives の数に対して true positives の数をプロットした ROC カーブ (receiver operating characteristics curve) [20] を生成した．ROC カーブでは，45 度の対角線はランダムな予測精度に相当し，左上に行けば行くほど，true positives の割合が増え，予測精度が良いことを表し，対角線に近づくようだと予測精度は悪いことを表す．表 3.1 で示した全ての実験に対して，この手順で ROC カーブを生成した．図 3.5 は，その結果を示す．単一のデータのみにカーネル主成分分析を適用した場合に比べると，検出率は，カーネル正準相関分析を用いたデータ比較やデータ融合を行うことによって向上していることが分かる．表 3.2 は，例として，閾値を全ての遺伝子の 10% としたとき (740 遺伝子から 74 個選んだ)，各手法に対して正しく検出されたオペロン遺伝子の数を表す.

オペロン遺伝子の検出法として，一番精度の良かった方法およびデータの組合せは，IKCCA-b であった．これは，*genome* と *pathway+expression* に適用したもの，つまり，“ゲノム上での並び” と，“パスウェイ+発現” に対する相関ということになる．次に良かったものは，IKCCA-a で，これは，*pathway* と *genome+expression* に適用したもの，つまり，“パスウェイ” と” ゲノム上の並び+発現 “の相関を表す.

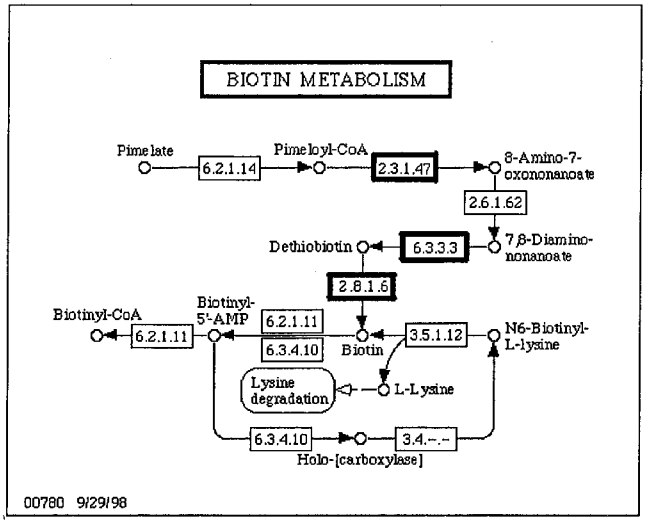

図 3.6: オペロンデータベースに登録されている既知のオペロンの例：
オペロンに属する遺伝子は，対応する EC 番号を実線で示されている．

性能が悪かったものは，OKCCA-c, IKCCA-c や OKCCA-b で，これは，*expression* とそれ以外のデータ (*pathway* や *genome*) との組合せにおいて実行されたものであった．

実際に検出した遺伝子を視覚化するため，KEGG/PATHWAY データベースへのマッピングを行った．ここでは，第 1 正準スコアの 10% を閾値として遺伝子群を選択した．一例として，ここでは，biotin metabolism に注目した．図 3.6 は，biotin metabolism に関わることが既知であるオペロンを表しており，それに属する 3 つの遺伝子群は実線で示されている．図 3.7 は， IKCCA-a 法によって選択された遺伝子を表しており，灰色で示されている．ここで，選択された遺伝子は，ほとんどオペロンに対応していることが観測された．

図 3.8 は，IKCCA-a 法によって選択された遺伝子の，実際のゲノム上での位置関係を示したものである．ここでは，遺伝子 JW0757, JW0758, JW0759, JW0761 は，パスウェイ上での遺伝子産物である酵素タンパク質の EC 番号 (Enzyme Comission Number) EC:2.6.1.62, EC:2.8.1.6, EC:2.3.1.47, EC:6.3.3.3 にそれぞれ対応していることに注意されたい．ただ遺伝子 JW0757 (EC:2.6.1.62) だけは，今回の採用したオペロンデータベース [25, 67] には，含まれていなかった．これらの検出された 4 つの遺伝子は， biotin metabolism 上で連続した化学反応を触媒し，それらはゲノム上で隣接して位置しており，ゲノム上で遺伝子クラスターを構成している．しかしながら，JW0757 の遺伝子の方向が，他のオペロン遺伝子と異なっている．この違いが，この遺伝子がオペロンデータベースに存在しない理由と考えること

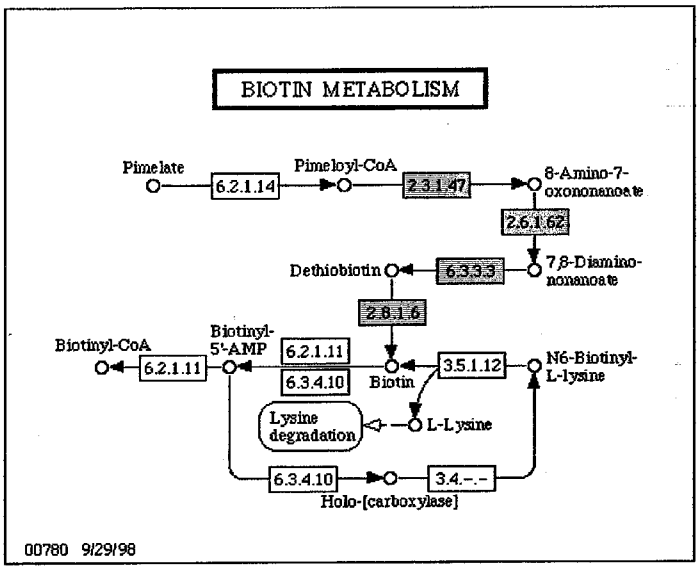

図 3.7: IKCCA で抽出されたオペロンの例：
選択された遺伝子は，対応する EC 番号を灰色で色付けされている．

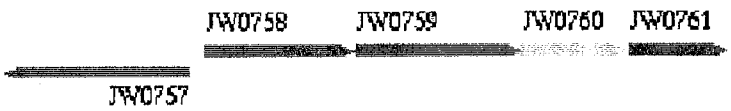

図 3.8: ゲノム上におけるオペロン遺伝子の例：
JW0757, JW0758, JW0759, JW0761 は，パスウェイ上の EC:2.6.1.62, EC:2.8.1.6, EC:2.3.1.47, EC:6.3.3.3 にそれぞれ対応する．

ができる．実際に，転写のメカニズムにおいては，遺伝子の向きは重要な因子である．なぜなら，遺伝子の上流のプロモータから同じ向きで，転写が開始するからである．それゆえ，このような遺伝子の向きなどの更なる生物学的な情報を反映させることによって，ここで提案している手法を更に改善することができると考えれる．

3.4 考察

本研究では，カーネル正準相関分析を一般化し，複数の異なるゲノムデータセット間の相関を解析する手法を開発した [63]．この手法の独自性として，数値ベクトルでないデータにも，相関という概念を拡張したことがあげられる．通常の多変量解析などの統計手法では，データが数値ベクトルでしか扱えない手法がほとんどだが，グラフや文字列などのどんな形式のデータ間の相関でも検出できるようにしたのが特長である．生化学パスウェイや，ゲノム配列，発現データなどの，複数のゲノムデータの比較解析や融合解析は，近年のバイオインフォマティクスの重要課題であった．データ構造が異なるとき，それらのデータの融合の手段としては，ヒューリスティックなアルゴリズムを採用するしか今まで手段が無く，それさえもデータ構造に依存する部分が大きかった．本研究では，様々な構造のデータを一つの統一的な枠組で解析できる方法を確立し，相関解析へ結びつけることができた．実際に，提案した手法を大腸菌のオペロン構造の検出法として適用したところ，オペロンに属する遺伝子群を上手く抽出することができ，複数の異質なゲノムデータ間に潜んでいる生物学的な相関を検出する方法としての有効性を示した．

オペロン遺伝子の検出ということに焦点を絞った結果から言えば，一番精度の良かった方法は，本研究で提案するデータ融合を組み合わせたカーネル正準相関分析 IKCCA であった．一番良かったデータの組み合わせは，IKCCA-b で，*genome* と *pathway* + *expression* に適用したもの，つまり，“ゲノム上での並び” と，“パスウェイ+発現” に対する相関ということになる．比較的性能が悪かったものは，*expression* とそれ以外のデータの組み合わせにおいて実行されたものであった．

これらの結果より，以下のような考察ができた．一つ目に，３つの生物学的属性間において，明確な階層性が見られた．*genome* は，オペロンに対して最も情報を多く含んでいる属性であり，*genome* 単一と他のデータを比較した結果が一番性能が良かったことからいえる．次に，*pathway* が重要な属性であると言える．*expression* は，本研究における解析では，オペロンとの関連が強く判断されにくかったデータであった．これは，数値結果からも，*expression* によって得られた正準相関を用いた実験で一番性能が悪かったことから推測できる．オペロン検出

において *genome* が一番寄与していたという結果は，オペロンは，ゲノム上における遺伝子のクラスターであるという，オペロンの本来の定義からも意味をなす．性能が悪かった OKCCA-b (*genome* vs. *expression*) や，*expression* 単一だけで他のデータと比較を行った他の実験結果からも言えることであるが，本研究で用いた遺伝子発現データの質があまり良くなかったという可能性もある．本来，オペロンに属する遺伝子は，共発現すると考えられるためである．対照的に，*pathway* の性能が良いのは，パスウェイデータベースの質が良かったとも言えるだろう．

二つ目として，データの質が良くないと考えられる発現データにも関わらず，一番良い結果は，*genome* と *pathway* + *expression* の比較から得られたものであった．これは，IKCCA-b は，*pathway* と *expression* の融合から，遺伝子発現データのノイズを除去し，*genome* と上手く相関する意味のある情報を抽出しているものと考えられる．実際に，*pathway* vs. *genome*，*expression* vs. *genome* という組合せよりも，性能が良かった．この実験結果は，各データセットに含まれているオペロンに対する情報の違いのために，IKCCA が，MKCCA や OKCCA よりも良い方法であるということを示す典型的な一例であると言えるであろう．

本研究のアルゴリズムでは，それぞれのデータセットをカーネル行列という遺伝子間の類似度行列に，全て変換している．これによって，複数の異質なデータセットを，一つの統一的な枠組で，数学的に扱えるようにしていることが特徴である．それゆえ，実際の性能は，どのカーネル関数を使ってデータを変換するか? または，どのように生物学的な知識を反映させた遺伝子間の類似度を設計するか? に依存する．そのため，様々なデータ構造に適したカーネル関数や，生物学的知識をカーネル関数に反映させるための方法論の開発が，近年のバイオインフォマティクスの分野で盛んに行われている [52, 55]．また，カーネル法とは，遺伝子間の類似度を要素とするカーネル行列を入力として，実用的な統計解析法，例えば，回帰分析や，判別分析，クラスタリング，主成分分析などを行おうという新しい統計学の手法であり，様々な目的に応じたデータ解析を行なうことができる [35, 46]．また，カーネル行列は一種の類似度行列を表すので，カーネルへの変換ができれば，本章で示したような異質なデータの融合も，和をとることで簡単に行なえるという長所もある．その意味で，異なるゲノムデータ間で，あたかも一種の共通言語のようにデータを取り扱えるという点が長所であり，今後のバイオインフォマティクスにおいて，標準的なゲノムデータ解析法となる可能性が高い [47]．

本研究におけるモチベーションは，先行研究でいえば，グラフ比較法に似ている [38, 37]．グラフ比較法では，複数のゲノムデータが与えられたとき，それを全てグラフに書き直す．そしてグラフ理論の観点から，相関遺伝子クラスターを見つける問題を，部分グラフイソモーフィズム問題に置き換えて，探索を行っている．しかしながら，グラフ比較法は，全ての遺伝子間の関係を 1 または 0 で置き換

えており，遺伝子間の関連の強さを考えてはいない．オペロン解析という目的に焦点を絞ったとき，グラフ比較法による相関クラスターとの理論的な関係，性能比較などは今後の課題である．

ここで提案した方法は，異質なゲノムデータセットが与えられたときに，それらの間にある相関構造をモデル化することを目的として開発した．単に，データ間の相関解析としても利用できる有効性を示したわけであるが，最終的には，高次元の生物学的機能を表すパスウェイ (タンパク質ネットワーク) を予測するための第一段階と考えて，この方法の開発を行なった．パスウェイデータは，他のゲノムワイドなデータ (遺伝子発現データや，酵母2ハイブリッド，DNA配列，系統プロファイル) に比べると，圧倒的に量が少ない．つまり，タンパク質間の機能的な関係に関しては，未知の部分が圧倒的に多い．それゆえ，このようなゲノムデータから，未知のパスウェイを予測しようする動きが急激に高まってきている．これは，もし未知のタンパク質間の機能的な関係を予測できれば，新しい生物学的な発見に直結するからである．次章では，本章で提案した相関解析法のモデルを用いて，未知のタンパク質間ネットワークを予測する方法を提案する．データとして，第2章において解析を行った系統プロファイルを一つのソースと考え，IKCCA のモデルを用いながら，未知のパスウェイの予測を試みる．

第4章　複数のゲノムデータからのタンパク質ネットワーク予測

4.1　序論

ゲノム情報から遺伝子やタンパク質によって成り立つ生命のはたらきを明らかにすることが，ゲノム解析の最終的な目的の一つである．生命のはたらきとは個々の遺伝子あるいはタンパク質に帰するものではなく，多数の遺伝子あるいはタンパク質が複雑に相互作用したネットワークのシステムで実現されるものである．その意味で，制御および代謝経路などのタンパク質のネットワークは生命システムの一部を表すため，ゲノム情報から未知のタンパク質のネットワークを予測することは，新しい生物学的な発見に直結するため意義がある．

生物工学の進歩によって，遺伝子やタンパク質に関するゲノムワイドなデータが蓄積されてきた．例えば，マイクロアレイ遺伝子発現データ [14, 49], 酵母2ハイブリッドによるタンパク質間相互作用情報 [53, 26], タンパク質の局在情報 [22], 系統プロファイル [43], パスウェイ情報 [28, 29] などが挙げられる．そこで，これらのゲノムデータや実験データを有効に使って，高次の生物学的な機能を表すタンパク質ネットワークを予測することが，近年のバイオインフォマティクスにおいて重要課題になっている．

本研究では，様々なゲノム情報から，生命システムを表すタンパク質ネットワークを予測する手法を開発した [64]．前章において異質なデータ間の相関解析を可能にしたカーネル正準相関分析を用いて，ゲノムデータとタンパク質ネットワークの相関モデルを構築し，新規のタンパク質間ネットワークを予測する方法を提案した．この方法の独自性は，教師付き学習の枠組においてネットワーク推定を行なう点にある．ここでいう教師付きとは，これまでに分かっている既知のタンパク質ネットワークの情報を予測過程の中で用いることを意味する．まず，第一段階として，ネットワークが既知のタンパク質セットから，ゲノムデータとパスウェイの相関 (ネットワーク構築原理) を，数学的に学習させ，モデルを構築する．第二段階として，そのモデルを，ネットワークの分かっていないタンパク質セットに当てはめ，ネットワークを予測する．教師付き学習の概念自体は，フィッシャーの判別分析，決定木，サポートベクターマシンなど，“個々のタンパク質の機能”

を予測を目的とする手法として先行研究でたくさんあるが，“タンパク質間の機能的関係”で構成されるネットワークを推定する手法は開発されておらず，本研究が最初の報告となる．ここでは，前章のゲノムデータの相関解析で用いた，カーネル正準相関分析のモデルを，ゲノムデータとパスウェイデータの相関関係を学習するのに用いる．

実際の適用例として，出芽酵母 *Saccharomyces cerevisiae* のタンパク質間の機能ネットワークを，マイクロアレイ遺伝子発現情報，酵母2ハイブリッドシステムによる相互作用情報，タンパク質の細胞内局在情報，系統プロファイルの4種類のデータから予測した．実験によって判明している既知のタンパク質ネットワークを用いて評価した結果，本研究で提案する複数のデータの統合と教師付き学習の効果によって，先行研究の方法（教師なし学習）よりも予測精度が著しく向上することが確認できた．そこで，全てのタンパク質セットに対して提案手法を適用し，出芽酵母の6059個のタンパク質からなる機能的ネットワークを推定した．それを基に，未知のタンパク質の機能や，missing 酵素の遺伝子候補を予測し，その妥当性について検討し，この手法が新しい生物学的な発見に繋がる可能性について議論した．

もう一つの適用例として，緑膿菌 *Pseudomonas aeruginosa* のリジン分解系におけるタンパク質ネットワークの再構築を試みた．ここでは，染色体上での遺伝子間の近さ，系統プロファイルによるタンパク質間の進化的な類似度を用いて，タンパク質の機能ネットワークを推定し，リジン分解系のパスウェイ上にあると思われる酵素遺伝子を予測した．EC:1.2.1.20，EC:2.6.1.48 などに対応すると予測された遺伝子について，大腸菌を宿主とした発現系を構築し酵素活性を確認したところ，実際に活性を示し，予測結果の妥当性を示唆した．

4.2 データ

4.2.1 タンパク質ネットワークの正解データ

出芽酵母 *Saccharomyces cerevisiae* のタンパク質ネットワークの正解データとして，KEGG/PATHWAY データベース [29] で保存されているタンパク質ネットワークを利用する．KEGG/PATHWAY データベースでは，タンパク質ネットワークは，頂点 (ノード) はタンパク質 (またはそれをコードする遺伝子)，辺 (エッジ) が以下に示す3種類のタンパク質間の機能的関係で構成される．一つはパスウェイにおいて連続的に化学反応を触媒する酵素間の関係，二つめは，タンパク質間の物理的相互作用の関係，三つめは転写因子とターゲットの遺伝子産物との遺伝子制御の関係である．ここでは，主に代謝パスウェイを表すタンパク質間の機能ネッ

トワークに注目した．つまり，連続的に化学反応を触媒する酵素間の関係で構成されるタンパク質ネットワークを考えた．最終的に，769 個のノード，3702 個のエッジから構成されるタンパク質ネットワークを作成した．以下では，これを信頼できるタンパク質ネットワークと見なし，後で提案するネットワーク予測法の性能を評価するための正解データとして扱う．

同様に，緑膿菌 *Pseudomonas aeruginosa* のタンパク質ネットワークの正解データとして，代謝パスウェイを表すタンパク質ネットワークを考えた．最終的に，799 個のノード，4472 個のエッジから構成されるタンパク質ネットワークを作成した．以下では，これを信頼できるタンパク質ネットワークと見なし，実際の適用例でで，missing 酵素遺伝子を予測するためのトレーニングデータとして用いる．

4.2.2 マイクロアレイ遺伝子発現データ

出芽酵母の遺伝子発現データは，Spellman らによる 77 種類の実験 [49]，Eisen らによる 80 種類の実験データ [14] を合わせた 157 種類の実験に基づくデータを用いた．各タンパク質をコードする遺伝子が，それぞれ 157 次元の数値ベクトルを持つデータセットとなる．

4.2.3 酵母 2 ハイブリッドシステム

2 種類の酵母 2 ハイブリッドの実験 [26, 53] に基づく，出芽酵母の 5470 個のタンパク質間物理的相互作用を用いた．酵母 2 ハイブリッドシステムは，多数の疑陽性 (false positives) の結果を示すことが問題点として指摘されており，ノイズの多いタンパク質間の関係を表すデータとみなすことができる．

4.2.4 タンパク質局在データ

出芽酵母のタンパク質の局在データは，網羅的に細胞内局在情報を調べた実験結果 [22, 71] から得た．このデータセットは，酵母の約 6234 個のタンパク質に対して，ゴルジ体，細胞質，小胞体，核内などの 23 個の細胞内局在のうち，出芽酵母のタンパク質が，どこで働いているかという情報を得ることができる．細胞内局在の例としては，例えば，ミトコンドリア，ゴルジ体，小胞体などがあげられる．各タンパク質に対して，局在プロファイルは，タンパク質は，ある局在に対して観察されれば，1，観察されなければ，0 で表される文字列である．

4.2.5 系統プロファイル

出芽酵母の系統プロファイルは，KEGGデータベースのオーソログクラスター[29]を基に作成した．この研究では，全ゲノム配列が解読されている，11種類の真核生物，16種類の古細菌，118種類の真正細菌の合計145生物種から構成される系統プロファイルとなる．ここでの系統プロファイルは，出芽酵母の各タンパク質をコードする遺伝子が，上の生物種に対して存在すれば1，存在しなければ0がコードされる文字列である．

同様に，緑膿菌の系統プロファイルも作成した．ここでの系統プロファイルは，緑膿菌の各タンパク質をコードする遺伝子が，上の生物種に対して存在すれば1，存在しなければ0がコードされる文字列である．

4.2.6 ゲノム上での位置情報

緑膿菌 *Pseudomonas aeruginosa* のタンパク質をコードする遺伝子のゲノム上での位置情報として，KEGG/GENESデータベース[29]に記述されている位置情報を利用する．染色体上での，遺伝子間の塩基数を距離と見なし，遺伝子間の近さを評価するのに用いる．

4.3 方法

4.3.1 カーネルによるデータ表現と統合

データ表現 ゲノムデータを統一的に計算機上で扱うため，全てのデータをカーネル行列[46]と呼ばれる類似度行列に変換することを提案する．直感的には，カーネルとは，あるデータセットに関して，タンパク質間の類似度またはタンパク質をコードする遺伝子間の類似度を表すものだと解釈できる．

例えば，データセットが，遺伝子発現データ，局在情報，系統プロファイルとすれば，ガウシアンカーネル $k(\mathbf{x},\mathbf{y}) = \exp(-\parallel \mathbf{x}-\mathbf{y} \parallel^2 /\sigma^2)$ や，線形カーネル $k(\mathbf{x},\mathbf{y}) = \mathbf{x}\cdot\mathbf{y}$ が自然な候補であろう．データが，タンパク質ネットワークや，酵母2ハイブリッドの相互作用などのグラフのときは，拡散カーネル[30]でカーネルに変換できる．

全てのデータをカーネルに変換する意義は，それぞれのデータ構造が，ベクトル，グラフ，文字列と異なっていたとしても，同じ数学的な枠組でデータを扱えるというメリットがある．

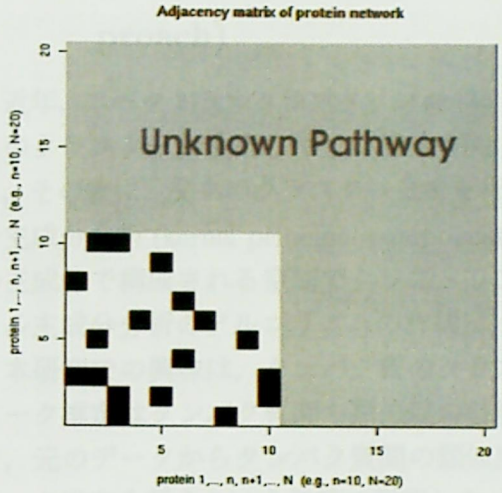

図 4.1: タンパク質ネットワークにおけるタンパク質の隣接行列の例

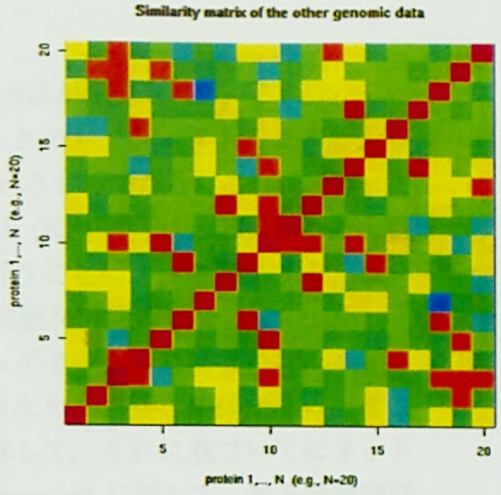

図 4.2: ゲノムデータに基づいて計算されたタンパク質の類似度行列の例

データの統合 ここで，$P \geq 1$ 個の異質なゲノムデータが得られており，それぞれ P 個のカーネル $K_1, \ldots, K_P$ で表されているとする．K_p は p 番目のデータセットに関する，タンパク質間の類似度行列を表す．一つの統合法として，第3章で説明した統合法を採用し，$K = \sum_{p=1}^{P} K_p$ と和を取ることでデータの統合をすることを提案する．

4.3.2 直接的なネットワーク推定法 (direct approach)

ここでは，複数のゲノムデータから，出芽酵母 *S. cerevisiae* のタンパク質ネットワークを予測することを考える．最初の，直接的な方法として，機能的に関連のあるタンパク質ペアは，データに関して高い類似度を持つという仮定をする．二つのタンパク質 $\mathbf{x}$ と $\mathbf{y}$ の類似度であるカーネルの値 $K(\mathbf{x}, \mathbf{y})$ が，ある閾値よりも大きければ，その2つのタンパク質ペアは相互作用するとみなす．それぞれのデータに関して，共発現するタンパク質ペア（タンパク質をコードする遺伝子ペア），物理的に相互作用するタンパク質相互作用ペア，同じ場所で働くタンパク質ペア，共進化するタンパク質ペア，それら情報を統合したものに基づき，タンパク質ネットワークを予測する．この離散バージョンは，グラフのジョイント法によるタンパク質間相互作用予測法 [34] に相当する．

4.3.3 教師なし学習に基づくネットワーク推定法 (spectral approach)

近年，スペクトラルクラスタリング [36] という方法が開発された．これは，データのクラスターが検出しやすい特徴空間に，データのオブジェクトをまず射影して，その後に，従来のクラスター分析を行おうというものである．これは，カーネル主成分分析 (kernel principal component analysis (KPCA)) [45] で得られる小数の主成分で構成される空間でクラスタリングを行うことに，ほぼ対応する．カーネル主成分分析のアルゴリズムの詳細は，参考文献 [45] を参照されたい．

本研究での興味は，タンパク質のクラスタリングそのものではないが，ネットワーク推定はタンパク質間の類似度の計算を伴うため，密接な関係がある．そこで，元のデータからタンパク質間の類似度を計算して，それに基づいてにネットワーク推定を行うという前節で説明した direct approach に対して，ある特徴空間に射影して，そこでタンパク質間の類似度を計算し，ネットワーク推定を行う方法が考えれる．ここでは，それを spectral approach と呼ぶことにする．

簡単に手順を説明すると，まず，各タンパク質 $\mathbf{x}$ を，ある特徴空間におけるベクトル $f(\mathbf{x}) = \left(f^{(1)}(\mathbf{x}), \ldots, f^{(L)}(\mathbf{x})\right)^{\mathsf{T}}$ に射影することを考える．ここで，$L < N$ であり，$f^{(l)}(\mathbf{x})$ は，l 番目の主成分に相当する．射影された特徴空間において，もう一度タンパク質間の類似度を計算し直し，再計算されたタンパク質間の類似度を基に，前節で述べた direct approach を実行する．次節で提案する教師付き学習に基づくネットワーク予測法と対比させると，これは，教師なし学習に基づくネットワーク推定法に対応する．

4.3.4 教師付き学習に基づくネットワーク推定法 (supervised approach)

実際に，我々が直面している状況を示したのが，図 4.1 と図 4.2 である．図 4.2 は，網羅的に得られたゲノムデータや実験データに基づくタンパク質間の類似度行列を表す．このようなゲノムデータから，高次の生物学的機能を表すタンパク質ネットワークを予測しようというのが目的である．図 4.1 は，タンパク質ネットワークの隣接行列を表す．ここで，黒色は，そのタンパク質間ペアは相互作用が存在する，白色は，その部分はタンパク質間相互作用しない (または確認されていない)，灰色は，その部分のタンパク質間の機能的な関係は未知である事を示す．ここでは，$n < N$ 個のタンパク質のネットワークは既知であり，N はタンパク質の総数を示す．

ここで，タンパク質ネットワークの情報の一部に関しては，得ることができる

ことに注意したい．つまり，ゲノムデータとタンパク質ネットワークの対応関係に関する知識を，一部のタンパク質セットに対しては得ることができるわけである．ここで，ゲノムデータからタンパク質ネットワークができる構築原理を，何らかの形で学習できれば，その構築原理を表すモデルを，ネットワーク情報が未知のタンパク質のセットに対して当てはめ，その部分のタンパク質間の相互作用の関係を予測できるのではないか？と考えた．前節で示した，direct approach と spectral approach は基本的に教師なし学習なので，その意味で，図 4.1 に示されているような事前知識を予測に用いておらず，図 4.2 に示されたゲノムデータだけを用いて，タンパク質ネットワークを探索的に予測していることに注意されたい．

本研究では，教師付き学習の枠組で，ゲノムデータとネットワークの事前知識の両方を用いて，タンパク質ネットワークを推定することを提案する．前節で述べた spectral approach を，教師付き学習になるように修正する．まず，各タンパク質 $\mathbf{x}$ を，ある特徴空間におけるベクトル $f(\mathbf{x}) = \left(f^{(1)}(\mathbf{x}), \ldots, f^{(L)}(\mathbf{x})\right)^{\top}$ に射影することを考える．ここで，$L < N$ であり，spectral approach では，$f^{(l)}(\mathbf{x})$ は，l 番目の成分に相当する．この射影の目的は，相互作用するタンパク質が，近くにいるような特徴空間を定義することである．それゆえ，$\mathbf{x}_i$ が $\mathbf{x}_j$ と相互作用するときは，$f(\mathbf{x}_i)$ は，$f(\mathbf{x}_j)$ と同じような特徴量であってほしいわけである．理想的には，これは，$f^{(l)}(\mathbf{x}_i)$ が $f^{(l)}(\mathbf{x}_j)$ に，$l = 1, \ldots, L$ に対して近ければよい．逆に，理想の特徴空間とは，もしタンパク質ネットワークが事前に分かるのであれば，関数 $f^{(l)}$ $(l = 1, \ldots, L)$ で構成される部分空間であり，タンパク質ネットワーク上での隣接するノード間で滑らかに変化するものであると言える．そのグラフに基づく拡散カーネルに関連するノルム $||f||$ は，その滑らかさの度合いを定量化したものになる [56]．つまり，f が滑らかであればあるほど，$||f||$ の値は小さくなる．その結果として，もしタンパク質ネットワークが既知であると仮定すれば，理想の特徴空間とは，グラフの拡散カーネルで主成分分析したときの主成分で構成される特徴空間となる．

実際には，真のタンパク質ネットワークの全ての情報は事前には知ることはできないので，その理想的な特徴空間への射影は求めることはできない．しかしながら，部分的には真のネットワークの情報を知ることができるので，その部分的な既知のネットワークに適合するような理想的な特徴空間を構築し，spectral approach によって作られる特徴空間を改良することを提案する．ここでは，全てのタンパク質の数を N とすると，n 個のタンパク質 $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ がネットワーク情報が分かっているタンパク質のセットであり，残りの $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N\}$ がネットワーク情報が分かっておらず，推定すべきタンパク質のセットとする．ここで，K_1 をネットワーク情報が既知のタンパク質に関するゲノムデータに基づき計算されたカーネル，K_2 をネットワーク情報が既知のタンパク質ネットワークから計算さ

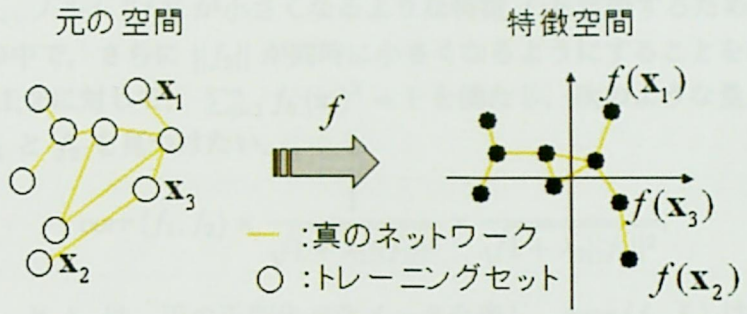

図 4.3: 教師付きネットワーク推定法 (supervised approach) のステップ1
ネットワーク情報が既知であるタンパク質をトレーニングデータセットとして用い，機能的な相互作用するタンパク質ペアが近くにあるような特徴空間を構築.

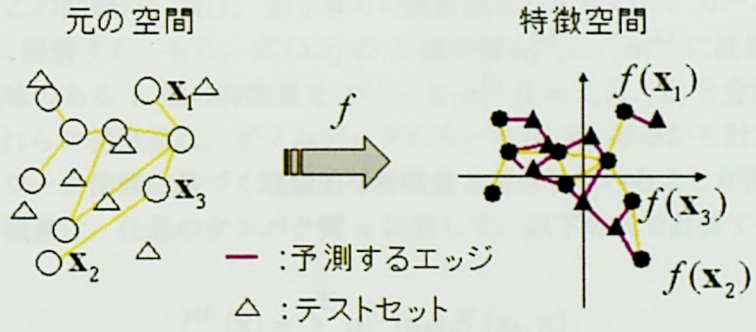

図 4.4: 教師付きネットワーク推定法 (supervised approach) のステップ2
相互作用が検出され易い特徴空間において，ネットワーク情報が未知であるテストセットのタンパク質の相互作用ペアを予測

れた拡散カーネルと定義する．K_1 と K_2 は両方とも $n \times n$ の行列であり，f を $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ に基づき定義された任意の関数，$||f_1||$ と $||f_2||$ をそれに対応するノルムとする．ノルム $||f_1||$ が小さくなるような特徴 f を定義するため，spectral approach の中で，さらに $||f_2||$ が同時に小さくなるようにすることを考える．ここで，$k = 1, 2$ に対して，$\sum_{i=1}^{n} f_k(\mathbf{x}_i)^2 = 1$ を満たし，次のような量を最大にするような f_1 と f_2 を見つけたい．

$$corr(f_1, f_2) \times \frac{1}{\sqrt{1 + \lambda_1 ||f_1||^2}} \times \frac{1}{\sqrt{1 + \lambda_2 ||f_2||^2}}, \tag{4.1}$$

ここで，λ_1 と λ_2 は，正の正則化パラメータを表し，$corr(f_1, f_2)$ は，f_1 と f_2 の標本相関係数を表す．この式の第一項は，f_1 の事前情報のネットワークに基づく f_2 への適合を示しており，第二項と第三項は $||f_1||$ と $||f_2||$ を小さく抑えることを意味している．このような特徴量は，逐次的に直交条件を追加し，同様の相関係数最大化の手順を適用することによって再帰的に求めることができる．この量 (4.1) を最小化することは，以下のような一般化固有値問題に帰着する．

$$\begin{pmatrix} \mathbf{0} & K_1 K_2 \\ K_2 K_1 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} (K_1 + \lambda_1 I)^2 & \mathbf{0} \\ \mathbf{0} & (K_2 + \lambda_2 I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}. \tag{4.2}$$

ここで，I は単位行列を示す．実際に，式 (4.1) の逐次的な解は $f_1 = K_1 \alpha_1$，$f_2 = K_2 \alpha_2$ と求めることができる．ここで，α_1 と α_2 は，式 (4.2) の固有ベクトルである．この問題は，実は，第3章での解析法として用いた，カーネル正準相関分析 [1] に帰着する．もし，式 (4.2) の L 個の解 $\alpha_1^{(1)}, \ldots, \alpha_1^{(L)}$ に注目するなら，それらは興味のある L 個の特徴量を $f^{(l)} = K_1 \alpha_1^{(l)}$ $(l = 1, \ldots, L)$ と定義することになる．これらの特徴量は，ゲノムデータのカーネル K_1 のみから計算でき，既知のネットワーク情報に基づく理想的な特徴量と適合していることが期待される．これらの特徴量は，任意のタンパク質 $\mathbf{x}$ に対して，以下の式で計算することができる．

$$f^{(l)}(\mathbf{x}) = \sum_{k=1}^{n} \alpha_1^{(l)}(\mathbf{x}_k) K(\mathbf{x}_k, \mathbf{x}). \tag{4.3}$$

この特徴量のセットが，タンパク質ネットワークを予測する前に射影を実行したときのタンパク質のセットである．

教師付きネットワーク推定法 (supervised approach) の視覚的なイメージを図4.3と図4.4にまとめた．第1段階として，ネットワーク情報が既知であるタンパク質セットをトレーニングセットとして用い，機能的な相互作用するタンパク質ペアが近くにあるような特徴空間を構築する．第2段階として，相互作用が検出され易い特徴空間において，ネットワーク情報が未知であるテストセットのタンパク

質の相互作用ペアを direct approach によって予測する．つまり，特徴空間において距離が近い (類似度が高い) タンパク質ペアにエッジを結ぶ．

spectral approach も supervised approach によって射影された各タンパク質 $\mathbf{x}$ は，L-次元のベクトルで，$\mathbf{u} = (u_1, \ldots, u_L)^\top = (f^{(1)}(\mathbf{x}), \ldots, f^{(L)}(\mathbf{x}))^\top$ と表される．射影後の特徴空間におけるタンパク質 $\mathbf{x}$ とタンパク質 $\mathbf{y}$ は，$\mathbf{u} = (u_1, \ldots, u_L)^\top$ and $\mathbf{v} = (v_1, \ldots, v_L)^\top$ と表され，そのネットワーク上におけるエッジとしての強さとして，ピアソンの相関係数のような以下の尺度を用いることにする．

$$\widehat{corr}(\mathbf{u}, \mathbf{v}) = \frac{\widehat{cov}(\mathbf{u}, \mathbf{v})}{\sqrt{\widehat{var}(\mathbf{u})}\sqrt{\widehat{var}(\mathbf{v})}} = \frac{\frac{1}{L}\sum_{l=1}^{L}(u_l - \bar{\mathbf{u}})(v_l - \bar{\mathbf{v}})}{\sqrt{\frac{1}{L}\sum_{l=1}^{L}(u_l - \bar{\mathbf{u}})^2}\sqrt{\frac{1}{L}\sum_{l=1}^{L}(v_l - \bar{\mathbf{v}})^2}}, \quad (4.4)$$

ここで，$\bar{\mathbf{u}}$ と $\bar{\mathbf{v}}$ は $\mathbf{u}$ と $\mathbf{v}$ の平均を表す．この値がある閾値よりも高ければ，タンパク質 $\mathbf{x}$ とタンパク質 $\mathbf{y}$ は，ネットワーク上で相互作用すると見なし，この値がある閾値よりも低ければ，ネットワーク上で相互作用しないであろうとみなす．この過程を全タンパク質ペアに行うことによって，網羅的なネットワークを予測する．

4.4 結果I：出芽酵母のタンパク質ネットワークの予測

4.4.1 ゲノムデータの変換

全てのゲノムデータを，まずカーネルに変換した．正解データのタンパク質間ネットワークと酵母2ハイブリッドのデータは，グラフ構造なので，拡散カーネルを用いて，K_{gold}， K_{y2h} とそれぞれカーネルの形に変換した．ここで，パラメータは $\beta = 1$ とした．遺伝子発現データは，実数値を値に取る数値ベクトルなので，ガウシアンカーネルを用いて K_{exp} と変換した．ここで，パラメータは，$\sigma = 5$ とした．局在データと系統プロファイルは，ビット列なので，線形カーネルを用いて K_{loc}， K_{phy} と変換した．最終的に，全てのカーネルは，対角成分が1になるように基準化した [46]．

結果として，遺伝子発現データ，酵母2ハイブリッド，タンパク質局在情報，系統プロファイル，正解のタンパク質ネットワークデータを，K_{exp}, K_{y2h}, K_{loc}, K_{phy}, K_{gold} と，それぞれカーネル行列に変換したことになる．

4.4.2 タンパク質間ネットワークの予測法としての性能評価

実際に，タンパク質間の機能的なネットワーク予測としての性能を見るため，direct approach と spectral approach の性能を，個々のゲノムデータと，全てのゲ

表 4.1: direct approach, spectral approach, supervised approach に対して行った数値実験の例

Approach	カーネル (データ)
Direct	K_{exp} (発現データ)
Direct	K_{y2h} (酵母 2 ハイブリッド)
Direct	K_{loc} (細胞内局在情報)
Direct	K_{phy} (系統プロファイル)
Direct	$K_{exp}+K_{y2h}+K_{loc}+K_{phy}$ (データ統合)

Approach	カーネル (データ)
Spectral	K_{exp} (発現データ)
Spectral	K_{y2h} (酵母 2 ハイブリッド)
Spectral	K_{loc} (細胞内局在情報)
Spectral	K_{phy} (系統プロファイル)
Spectral	$K_{exp}+K_{y2h}+K_{loc}+K_{phy}$ (データ統合)

Approach	カーネル (データ)	カーネル (ターゲット)
Supervised	K_{exp} (発現データ)	K_{gold} (タンパク質ネットワーク)
Supervised	K_{y2h} (酵母 2 ハイブリッド)	K_{gold} (タンパク質ネットワーク)
Supervised	K_{loc} (細胞内局在情報)	K_{gold} (タンパク質ネットワーク)
Supervised	K_{phy} (系統プロファイル)	K_{gold} (タンパク質ネットワーク)
Supervised	$K_{exp}+K_{y2h}+K_{loc}+K_{phy}$ (データ統合)	K_{gold} (タンパク質ネットワーク)

ノムデータを統合したカーネルの両方に対して適用した．全ての実行手順およびデータの組合せのリストを，表 4.1 の上段，中段に示す．spectral approach に対しては，最初の $L = 50$ 個の主成分を，特徴空間を構成するために用いた．予測精度は，正解データのタンパク質ネットワークをどれだけ復元できるかで評価した．ある閾値を設定し，タンパク質ペアの類似度が，閾値よりも大きい時そのタンパク質ペアは機能的な相互作用があると予測し，その閾値よりも小さい時そのタンパク質ペアは機能的な相互作用がないと予測する．閾値の値を，小さい値から少しずつ大きくしていき，それぞれの閾値の値でエッジの有無を予測したときの，true positives (予測したエッジが実際に正解データの中にあるとき) の数と， false positives (予測したエッジが正解データに無いとき) の数を記録していった．

図 4.5 と図 4.6 は，変化させていった閾値の値に対し，false positives の割合に対して true positives の割合をプロットした ROC カーブ [20] を示している．ROC カーブでは，45 度の対角線はランダムな予測精度に相当し，左上に行けば行くほど，true positives の割合が増え予測精度が良いことを表し，対角線に近づくようだと予測精度は悪いことを表す．両方の場合とも，45 度の対角線より少し上にプロットされているが，全体的な予測精度は，あまり良くないことが図から読み取れる．direct approach に比べると，spectral approach は，特にデータを統合したとき，少し精度が改善していることが分かる．これらの結果は，タンパク質ネットワークの推定問題は，非常に難しい問題ということを示している．

次に，教師付き学習に基づくネットワーク推定法を適用した．アルゴリズムの正則化パラメータ λ_1 と λ_2 はそれぞれ 0.1 とおき，特徴空間の次元数として，$L = 50$ 個の特徴量を使って特徴空間を構築した．また，個々のゲノムデータのタンパク質相互作用予測への重要性，全てのデータを統合した時の効果の両方を見るために，それぞれの場合に対して性能の検証を行った．全ての実行手順およびデータの組合せのリストを，表 4.1 の下段に示す．予測精度を測るために，以下のようなクロスバリデーション実験を行った．まず，769 個のタンパク質のセットを，9 対 1 の割合で，トレーニングデータとテストデータに，ランダムに分割する．次に，トレーニングデータを基に特徴空間を学習し，テストデータのタンパク質が持つ可能性のあるペア (図 4.1 の灰色部分) のタンパク質相互作用について予測を行った．

このクロスバリデーションの過程を 10 回行い，そこで生成される ROC カーブの平均をプロットしたのが，図 4.7 である．この教師付き学習のネットワーク推定法で，精度が向上していることが分かる．個々のゲノムデータの中では，発現データと系統プロファイルが，高い寄与を与えているので，代謝パスウェイ上の機能的なタンパク質間相互作用の予測には，遺伝子の共発現の情報とタンパク質の進化的な情報が重要であることがわかる．次にタンパク質の細胞内局在情報が重要

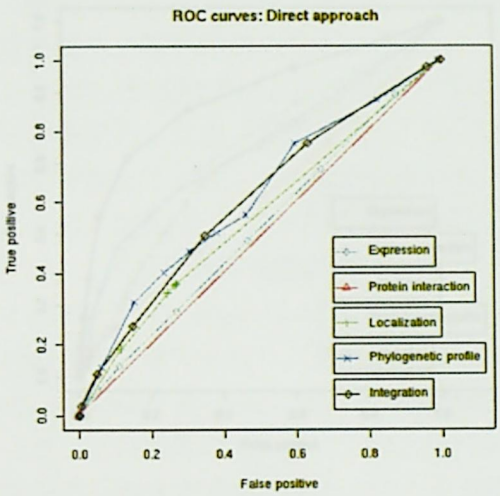

図 4.5: ROC カーブ: Direct approach
水色は，発現データのみを使った結果，赤色は，酵母 2 ハイブリッドのみを使った結果，緑色は，細胞内局在情報だけを使った結果，紺色は，系統プロファイルだけを使った結果，黒色は，全てのデータを統合した結果を表す．

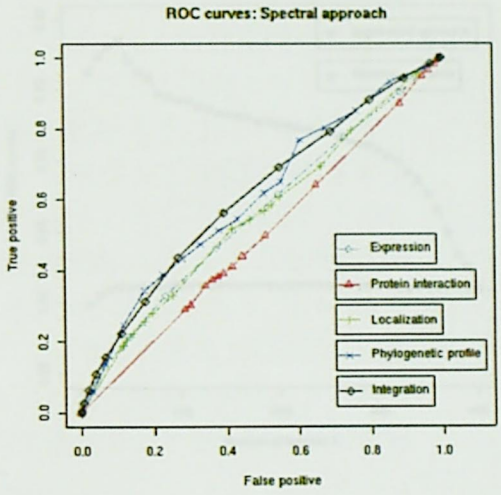

図 4.6: ROC カーブ: Spectral approach
水色は，発現データのみを使った結果，赤色は，酵母 2 ハイブリッドのみを使った結果，緑色は，細胞内局在情報だけを使った結果，紺色は，系統プロファイルだけを使った結果，黒色は，全てのデータを統合した結果を表す．

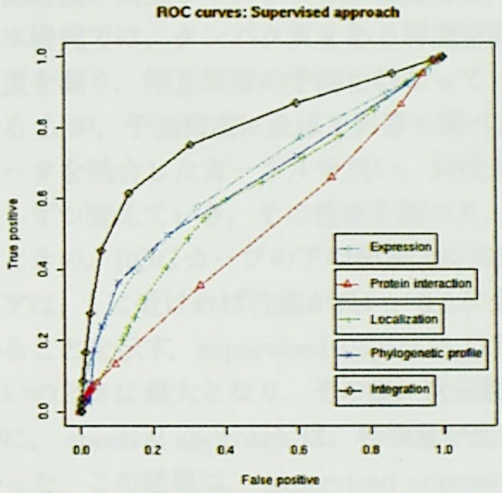

図 4.7: ROC カーブ: Supervised approach
水色は，発現データのみを使った結果，赤色は，酵母 2 ハイブリッドのみを使った結果，緑色は，細胞内局在情報だけを使った結果，紺色は，系統プロファイルだけを使った結果，黒色は，全てのデータを統合した結果を表す．

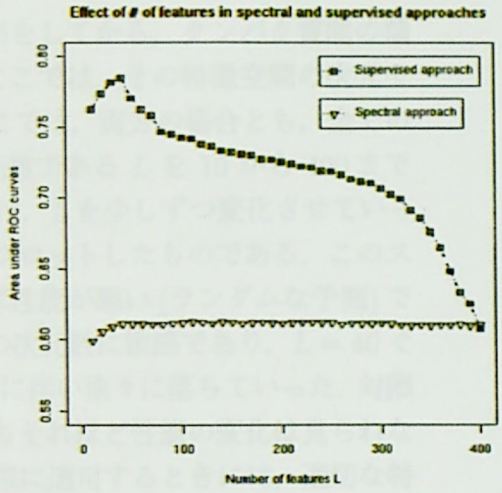

図 4.8: spectral approach, supervised approach における特徴量の個数の影響
特徴空間の次元数 L を，10 から 400 まで，少しずつ変化させていったときの，ROC カーブの下の面積の変化．x 軸は特徴空間の次元数，y 軸は ROC カーブ下の面積を表す．三角のマークは，spectral approach の結果を示し，菱形のマークは，supervised approach の結果を示す．

で，酵母2ハイブリッドのデータは代謝パスウェイ上のタンパク質間相互作用には，ほとんど関連は無いことを示唆している．このことは，酵母2ハイブリッドによる物理的なタンパク質相互作用ペアは，代謝ネットワークで化学反応を連続的に触媒する酵素間ペアと関係ないことは生物学的に明らかなので，妥当な結果であると言えよう．全ての結果を比較した結果，全てのゲノムデータの統合し，かつ教師付き学習に基づくネットワーク推定を行なった結果が，一番良いことが分かる．つまり，様々なゲノムデータの統合，教師付き学習の二つの効果によって，予測精度が向上していることが確認できた．

本研究では，タンパク質をある特徴空間に射影をしてから，タンパク質間の類似度を測り，相互作用の予測を行なっている．ここでは，その特徴空間の次元である L が，予測精度に及ぼす影響を調べた．ここでは，両方の場合とも，全てのデータを統合したカーネルを用い，特徴量の次元数である L を 10 から 400 まで少しずつ変えていき，その性能を調べた．図 4.8 は，L を少しずつ変化させていったときの，ROC カーブの下の面積 [20] の変化をプロットしたものである．このスコアは，1 に近ければ性能が良い，0.5 に近ければ性能が悪い (ランダムな予測) であることを示す．supervised approach は特徴量の次元数に敏感であり，$L = 40$ ぐらいのときに最大となり，その後，次元数の増加に伴い徐々に落ちていった．対照的に， spectral approach は，特徴量が変わってもそれほど性能の変化は見られなかった．この結果は， supervised approach を実際に適用するときには，適切な特徴量の数を設定する必要があることを示唆している．

4.4.3 全タンパク質に対する網羅的なネットワーク予測

クロスバリデーション実験によって，本研究で提案するネットワーク推定法の妥当性が確認できたので， 次に全タンパク質を使って網羅的なタンパク質ネットワークの予測を行った．ここでは，酵母2ハイブリッドのデータは使わず，遺伝子発現データ，タンパク質局在情報，系統プロファイルの3種類のデータから，出芽酵母の 6059 個の ORF をノードとするネットワークを予測した．この予測したタンパク質間ネットワークによって，様々な新しい生物学的な考察をすることが可能になることが考えられる．例えば，1）missing 酵素遺伝子の同定，2）機能未知タンパク質の機能予測，3）機能未知なタンパク質間相互作用の予測などが挙げられる．以下では，それらの例を示す．

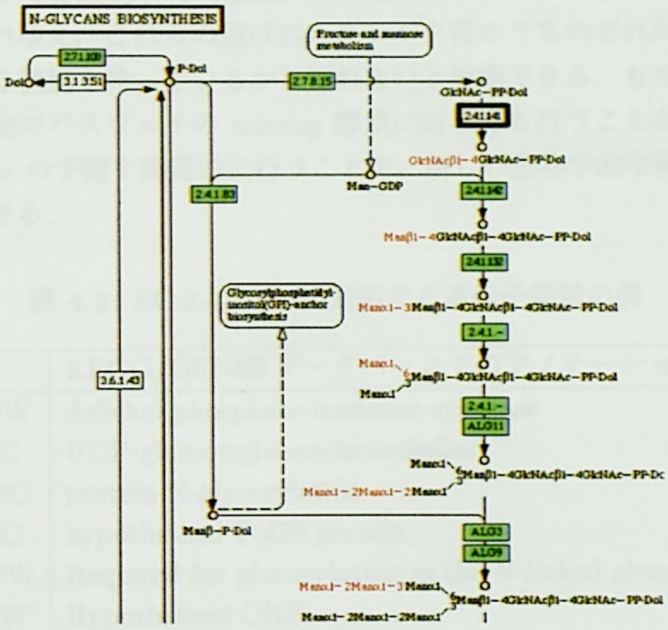

図 4.9: N-糖鎖生合成パスウェイの一部
EC:2.4.1.141 は missing 酵素に相当する.

4.4.4 missing 酵素遺伝子の同定

ある酵素が代謝パスウェイにあると分かっているが，その酵素遺伝子が同定されていないという missing 酵素遺伝子の同定は，生化学において重要な課題である．例えば，図 4.9 で示している N 型糖鎖生合成のパスウェイでは，EC:2.7.8.15 と EC:2.4.1.142 の間に，一つ missing 酵素 (EC:2.4.1.141) が存在する．閾値 0.6 で予測したときのタンパク質ネットワークを基に，EC:2.7.8.15 と EC:2.4.1.142 につながっている ORF を探索した．表 4.2 で，その missing 酵素の候補として自動的に予測された遺伝子の例を示している．ほとんどが糖鎖関連遺伝子であり，また YPL207W と YGL010W など機能が良く分かっていない遺伝子なども多数含まれていた．それゆえ，これらの選ばれたタンパク質のうちのどれかが，この化学反応を触媒する機能を持っているかも知れないと推測できる．もちろん，このような推論は，他のパスウェイの missing 酵素に対しても行うことができる．この missing enzyme の予測を網羅的に行うことで，新しい生物学的な発見につながることが期待できる．

表 4.2: EC:2.4.1.141 に対応する遺伝子候補の例

遺伝子	KEGG/GENES データベースでのアノテーション
YPR183W	dolichol phosphate mannose synthase
YPL227C	UDP-glucose:glucosyltransferase
YMR013C	protein N-glycosylation
YBL082C	hypothetical F-458 protein
YOR002W	Required for glucosylation in the N-linked glycosylation
YPL207W	Hypothetical ORF
YPR003C	Hypothetical ORF
YGL010W	Hypothetical ORF
YNL125C	similarity to mammalian monocarboxylate transporters
YOR285W	Hypothetical ORF
...	...

4.4.5 タンパク質の機能予測

次に，予測した網羅的なタンパク質ネットワークを用いて，機能未知であるタンパク質の生物学的機能を予測する例を示す．ここでは，機能がよく分かっていないタンパク質 YJR137C に注目した．2003 年 9 月の段階では，その酵素としての機

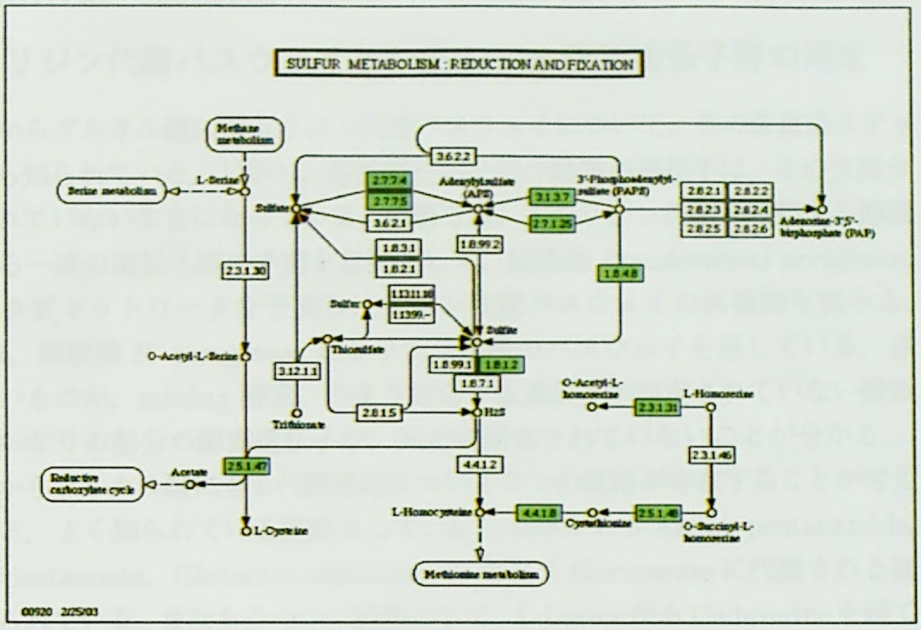

図 4.10: 硫黄の代謝パスウェイ

能は未知であり，EC 番号は分かっていなかった．つまり，学習するためのタンパク質ネットワークの既知ネットワークには，入っていなかったタンパク質である．我々の予測したネットワークを見てやると，YJR137C は，酵素 EC:1.8.4.8 と，酵素 EC:2.5.1.47 に繋がっていた．この二つの酵素 EC:1.8.4.8 と，酵素 EC:2.5.1.47 は，硫黄の代謝パスウェイで働くことが知られているので，この YJR137C は，硫黄に関連するような生物学的な機能を持つのではないかと推測できる．図 4.10 は，出芽酵母の硫黄の代謝パスウェイを示している．また，このパスウェイで，ターゲットのタンパク質 YJR137C は，酵素 EC:1.8.4.8 と，酵素 EC:2.5.1.47 に，連続して化学反応を触媒する機能があるのではないかと推測でき，KEGG/PATHWAY データベースのリファレンスパスウェイにある EC 番号 EC:1.8.1.2 に相当するのではないかと予測できる．

近年，出芽酵母のコミュニティデータベースである MIPS データベース [68] において，YJR137C は， EC 番号 EC:1.8.1.2 に対応する酵素であるという報告がされていた．つまり，予測したネットワークに基づた，このタンパク質の機能予測は当たっていたことを意味する．本研究で提案する教師付き学習に基づくネットワーク推定法の有効性を支持する結果といえるであろう．

4.5 結果II：緑膿菌のタンパク質ネットワークの予測

4.5.1 リジン代謝パスウェイ上のmissing 酵素遺伝子群の同定

リジンからグルタル酸に至るリジン代謝パスウェイについて，その各反応ステップ古くから知られている．しかし，各触媒反応を行う酵素の遺伝子は，その大部分が同定されていないままになっている．本節では，このリジン代謝を触媒する酵素に対応する一連の遺伝子群の予測を目的として，緑膿菌 *Pseudomonas aeruginosa* のタンパク質ネットワークを予測し，リジン代謝パスウェイの再構築を試みる．図 4.11 は，緑膿菌 *P. aeruginosa* のリジン分解系のパスウェイを示している．赤で表しているのが，missing 酵素，つまり対応する遺伝子が同定されていない酵素を示す．かなりの部分の酵素遺伝子が，未だに同定されていないことが分かる．

リジンからグルタル酸に至る代謝経路について２つの経路が存在することが考えられている．よく知られている経路として，L-Lysine から 5-Amino-pentanamide, 5-Amino-pentanoate, Glutarate semialdehyde を経て Gurutarate に代謝される経路が考えられている．またもう一つの経路として，L-Lysine から Cadaverine を経て 5-Amino-pentanoate に代謝されグルタル酸に至る Cadaverine 経路の存在が緑膿菌 *P. aeruginosa* において考えられている．そこで，本研究では，まずこれら二つの経路に共通する 5-Amino-pentanoate から Glutarate semialdehyde を経て Glutarate に至る二つの酵素遺伝子の同定に焦点を絞ることにする．

4.5.2 バクテリアゲノムの特徴を反映させたカーネル

緑膿菌 *P. aeruginosa* に関する網羅的な実験データはあまり無いので，ここでは，バクテリアゲノムの特徴を利用して，タンパク質ネットワークの推定を試みる．バクテリア遺伝子が持つ特徴として，以下の２つの特徴があることが言われている．

- 機能的に関連のあるタンパク質の遺伝子はゲノム上で近い位置にある傾向がある．[8]
- 機能的に関連のあるタンパク質は同じような進化パターンを持つ傾向がある．[43, 41]

これらの特徴を利用するため，ゲノム上での遺伝子の位置情報，系統プロファイルの２種類のデータを用いた．

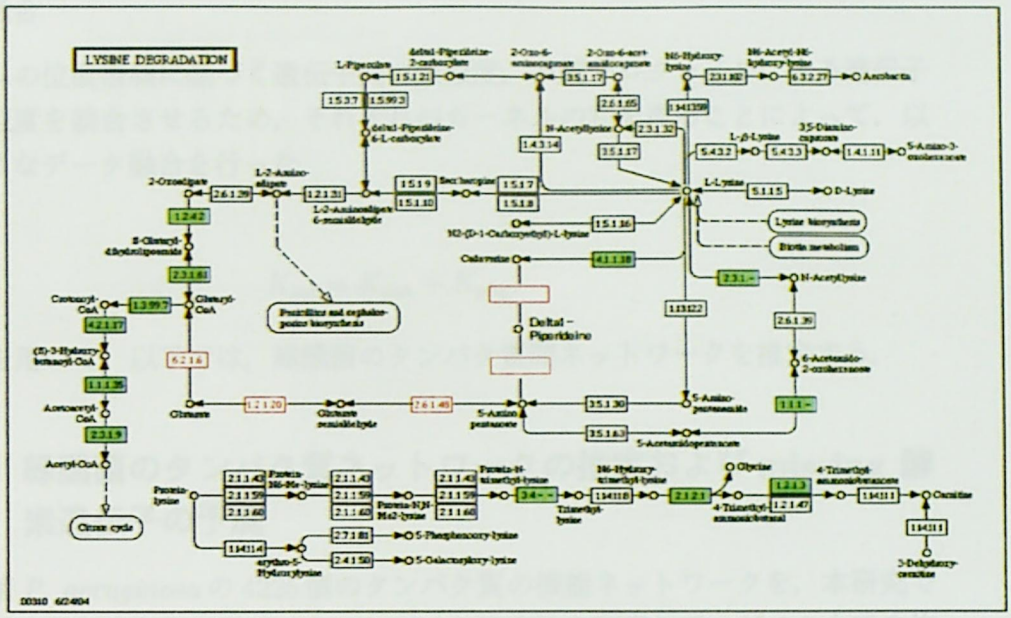

図 4.11: 緑膿菌のリジン分解系のパスウェイ

ゲノム上での位置情報

遺伝子間の距離は，染色体上での2つの遺伝子間の塩基数として考えることにする．遺伝子間距離が近ければ類似度が高くなり，遺伝子間距離が遠ければ類似度が低くなるようなスコアを構築するため，以下のようなカーネルを考えた．

$$K_{gen}(\mathbf{x}, \mathbf{x}') = \exp(-d/h).$$

ここで，d は，遺伝子 $\mathbf{x}$ と遺伝子 $\mathbf{x}'$ 遺伝子の距離 (塩基数) であり，h は，正のパラメータを表す．ここでは，$h = 10000$ とした．

系統プロファイル

ここでの系統プロファイルは，緑膿菌 *P. aeruginosa* の各タンパク質をコードする遺伝子が，145種の生物種に対して存在すれば1，存在しなければ0がコードされる文字列である．

線形カーネルを用いて，2つの遺伝子間の系統プロファイルの内積を計算し，遺伝子 $\mathbf{x}$ と遺伝子 $\mathbf{x}'$ の類似度を表すカーネルを以下のように計算した．

$$K_{phy}(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'.$$

ここで，$\mathbf{x}$ は，145次元のビット列となっている．

データ融合

ゲノムの位置情報に基づく遺伝子間の類似度，系統プロファイルによる遺伝子間の類似度を統合させるため，それぞれのカーネルの和を取ることによって，以下のようなデータ融合を行った．

$$K_{int} = K_{gen} + K_{phy}.$$

これを用いて，以下では，緑膿菌のタンパク質間ネットワークを推定する．

4.5.3 緑膿菌のタンパク質ネットワークの推定およびmissing 酵素遺伝子の予測

緑膿菌 *P. aeruginosa* の 4225 個のタンパク質の機能ネットワークを，本研究で提案した教師付き学習に基づくネットワーク推定法を利用して，ゲノム上での位置情報，系統プロファイルから予測した．予測したネットワークにおいて，リジン代謝パスウェイにおける既知遺伝子と，予測したネットワーク上において近い位置にある遺伝子を，ターゲットの候補遺伝子として予測した．その結果，リジン分解系で missing 酵素である EC:1.2.1.20 と EC:2.6.1.48 に対応する候補遺伝子として，PA0265 と PA0266 を予測した．表 4.3 は，今回注目しているリジン分解系パスウェイにおける missing 酵素を予測した遺伝子候補のリストを示している．

表 4.3: missing 酵素の遺伝子候補のリスト

missing 酵素	予測した遺伝子	KEGG/GENES におけるアノテーション
EC:6.2.1.6	PA0262, PA0260	Hypothetical protein
EC:1.2.1.20	PA0265	dehydrogenase (EC:1.2.1.16)
EC:2.6.1.48	PA0266	amino-transferase (EC:2.6.1.19)
Delta. → 5-amino.	PA1252	dehydrogenase (EC:1.1.1.13)
EC:3.5.1.30	作業中	作業中
EC:1.13.12.2	作業中	作業中
Cadav. → Delta.	作業中	作業中

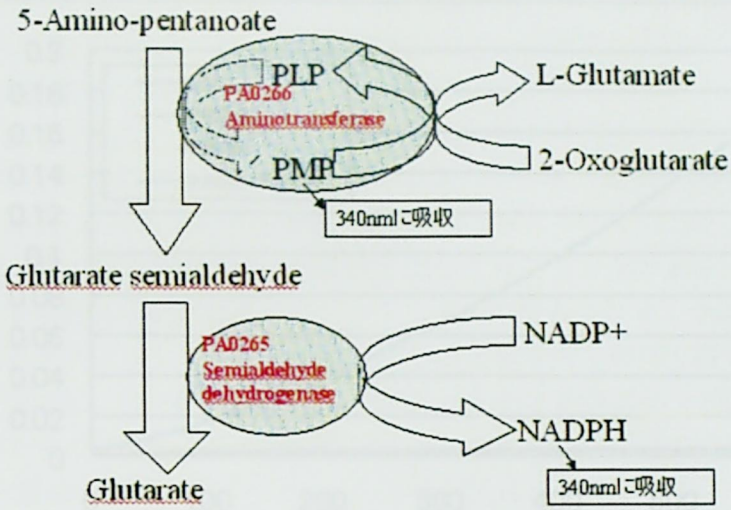

図 4.12: ターゲットの連続した化学反応

4.5.4 実験による検証

今回予測した酵素遺伝子のうち，PA0265，PA0266 について実験による検証を行った．PA0265，PA0266 についてクローニングを行い，大腸菌による発現系を構築した．これら遺伝子が予測した代謝反応を行うかどうかを in vitro における触媒反応実験によって確かめた．図 4.12 は，2 つの連続した化学反応における基質と生成物を表したものである．

我々が予測した 5-Amino-pentanoate から Glutarate semialdehyde を経て Glutarate に至る反応は二つの反応によって触媒される．まず，第一反応のアミノ基転移反応において 5-Amino-pentanoate から Glutarate semialdehyde が生成する．

5-Amino-pentanoate + 2-Oxoglutarate ⇔ Glutarate semialdehyde + L-Glutamate

さらに第二反応の脱水素反応によって Glutarate semialdehyde から Glutarate が生成する．

Glutarate semialdehyde + NAD+ + H2O ⇔ Glutarate + NADH + H+

これら二つの反応を我々が予測した遺伝子が触媒することを確認するために，予測遺伝子が 5-Amino-pentanoate，2-Oxoglutarate，NAD+ を基質として Glutarate

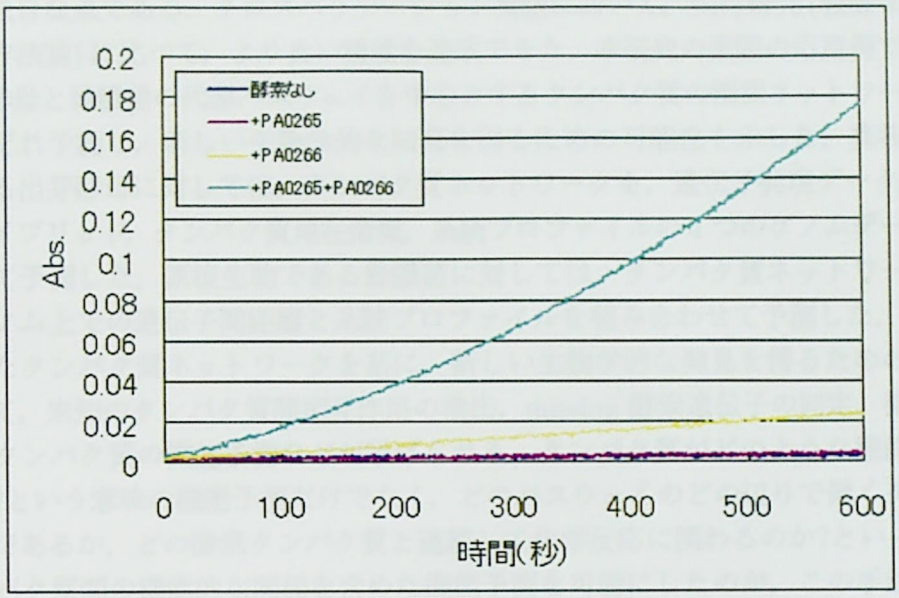

図 4.13: 吸光度の経時的変化

を生成することができるかどうかを第二反応で生じるNADPHの340nmにおける吸光の増大を測定することによって確認した．

図4.13は，5-Amino-pentanoate，2-Oxoglutarate，NAD+を基質として予測遺伝子，PA0265及びPA0266を加えたときの340nmの経時的な吸光度の変化を示している．予測遺伝子をそれぞれ一方だけ加えたものについては，ほとんど吸光の変化が見られず，共に加えた場合に大きな増大が見られた．これは第二反応の生成物であるNADPHによる吸光の増大であると考えられることから，我々が予測した遺伝子はパスウェイ上におけるこれら一連の反応を触媒することが示された．またPA0266のみを加えた場合にわずかに吸光の増大を確認することができた．これは第一反応で生成するPMPによる吸光と考えられるので，PA0266が第一反応を触媒することが示唆される．それゆえ，今回予測した遺伝子が，ターゲットのmissing酵素の遺伝子であることの妥当性を示した．

4.6 考察

この章では，高次の生物学的機能を表すタンパク質ネットワークを，複数のゲノムデータから予測する手法を提案した．提案手法では，カーネル正準相関分析のモデルを用い，教師付き学習の枠組みにおいて，ネットワーク推定を行なって

いる点が独自な点である．クロスバリデーション実験において，先行研究(教師なし学習の方法論)に比べて，より良い精度を達成できた．本研究の実際の応用例では，出芽酵母と緑膿菌の代謝パスウェイを中心とするタンパク質の機能ネットワークをそれぞれ予測し，新しい生物学的な知見を得るための可能性を示した．真核生物である出芽酵母に対しては，タンパク質ネットワークを，遺伝子発現データ，酵母２ハイブリッド，タンパク質局在情報，系統プロファイルの４つのゲノムデータを用いて予測した．原核生物である緑膿菌に対しては，タンパク質ネットワークを，ゲノム上での遺伝子間距離と系統プロファイルを組み合わせて予測した．

予測したタンパク質ネットワークを基に，新しい生物学的な発見を得るための応用として，未知のタンパク質間相互作用の検出，missing 酵素遺伝子の同定，機能未知のタンパク質の機能予測などが挙げられる．タンパク質がどのような機能を持つか?という意味の機能予測だけでなく，どのパスウェイのどの辺りで働くタンパク質であるか，どの酵素タンパク質と連続して化学反応に関わるのか?といった，タンパク質間の機能的な関係を含めた機能予測を可能にしたのが，この手法の特長である．例として，missing 酵素の遺伝子の同定，タンパク質の機能予測の例をあげたが，同じような予測は，他のパスウェイや他のタンパク質に対しても行うことができる．ただし，新しい生物学的な発見ができたという確証を得るには，例えば，missing 酵素遺伝子の候補として予測した遺伝子が実際に酵素活性を持つかどうか，実際に実験をして確認する必要があるだろう．

現在，著者は実験系の生物学者と共同研究を進め，本研究で提案するタンパク質ネットワーク予測法に基づく，網羅的な missing 酵素の遺伝子の同定の作業を行っているところである．今回の応用例の一つである，緑膿菌のリジン分解系のパスウェイにおける missing 酵素遺伝子の同定は，その一例である．今回は，同じファミリーの中から候補遺伝子が取れてきたが，1) 多数のファミリーの中から選ぶことができる，2) 配列類似性が全く無い場合でも検出ができる，という利点もある．これは，従来の配列類似性に頼った酵素遺伝子の同定法では不可能であり，本研究で提案する手法の独自な利点である．現段階では，今回対象としたリジン代謝パスウェイを完全に再構築しきれていないが，現在，計算機による予測および実験による検証を同時に進めている．

アルゴリズムの観点からみると，本研究で提案する手法は教師付き学習であるのに対し，今まで提案されている先行研究の手法は全て教師なし学習に属する．教師付き学習では，アルゴリズムの中で，既知のネットワークとそれに対応するゲノムデータの相関を自動的に学習できる点が特徴である．それゆえ，生化学的な代謝パスウェイに限らず，遺伝子制御ネットワークや，シグナリングパスウェイ，物理的なタンパク質間相互作用ネットワークなど，学習過程で使うターゲットのネットワークを替えるだけで，様々な種類のネットワーク推定に利用することが

できる．もう一つの長所として，異質なデータを同時に統合できるという点にある．データ構造に適したカーネル関数を使って，タンパク質間の類似度行列さえ変換できれば，どのようなデータでも統一的な枠組で扱うことができる．より最適なカーネル関数やそのパラメータの選択といった問題は，今後の課題である．

第5章　全体への結論

5.1　本研究のまとめ

生物科学は，ポストゲノム時代に突入し，大量のゲノム情報が溢れるようになった．これらの情報を表すデータを効率良く解析し，どうやって有益な解釈を得るかということが今後のバイオインフォマティクスの焦点である．

本研究では，一つのアプローチとして，カーネル法という統計手法のアイデアを採用し，ゲノム情報の異質性を統一的な枠組で解析できるような方法論を開発した．この方法によって，配列情報，発現情報，進化情報，相互作用情報などの，個々のゲノム情報の解析だけでなく，様々なゲノム情報間の相関を解析することを可能にした．実際に，原核生物の大腸菌の遺伝子に関する，パスウェイ，ゲノム上のでの並び，遺伝子発現データの3つのデータからオペロン構造を抽出するのに，その有効性を示した．

さらに，この相関解析法の数学モデルを応用し，新規のタンパク質ネットワークを予測する方法論の提案を行った．この方法により，未知のタンパク質間相互作用によって構成されるネットワークを予測できるだけでなく，missing 酵素遺伝子の同定や，未知のタンパク質の機能予測など，新しい生物学的な知見を得るための可能性を示した．

5.2　今後の展望

今後は，本研究で提案したネットワーク推定法によって予測した新規のタンパク質間相互作用ペアを，新しい生物学的な発見に繋げるために，実験系の生物学者と共同研究を行うことで，実験的な確認作業を行いたいと考えている．特に，代謝パスウェイに存在する多数の missing 酵素遺伝子の網羅的な同定を進める予定である．

今回は代謝ネットワークに注目したが，本研究で提案するネットワーク推定法を用いることで，遺伝子制御ネットワークなどの他の生物学的なネットワーク予測も行うことも可能であることを記しておきたい．ただし，遺伝子制御ネットワークでは，どの遺伝子がどの遺伝子を制御するのか?という方向の情報が大切なので，

そういった方向の情報も予測できるようにモデルの拡張を進めて行く予定である.

謝辞

本研究は，京都大学化学研究所バイオインフォマティクスセンターの生命知識システム領域(金久研究室)において，金久實教授の指導の下に行われました．同教授にはバイオインフォマティクスという新しい分野で研究する機会と環境と与えて頂くと共に，研究の方向性や学術論文の構成について，多くの助言を頂きました．深く感謝致します．

同研究室の五斗進助教授には，研究の進行状況の相談や論文の推敲など細部に至るまで，多くのご進言とご指導を与えて頂いた上，日常を通じて様々な面倒をみて頂きました．Ecole des Mines de Paris の Jean-Philippe Vert 氏には，本研究を進める上で必要な数学的な側面から，多くのご進言とご指導して頂きました．同じバイオインフォマティクスセンターの阿久津達也教授には，ご指導を頂いただけではなく，フランスと日本の共同研究のプロジェクトメンバーに入れて頂き，国際的な共同研究の機会を与えて頂きました．深く感謝致します．

同研究室の川島秀一助手，服部正泰助手，片山俊明助手には，セミナーや日常を通じて面倒をみて頂きました．吉沢明康氏，伊藤真純氏，檜作好之氏，佐藤哲也氏には，本研究における解析結果に対する生物学的な考察において，多くの助言をして頂くとともに，実際に解析作業を手伝って頂きました．大変有難うございました．

第4章の緑膿菌のリジン分解系における missing 酵素遺伝子の予測は，京都大学化学研究所生体分子機能研究部門 II の江崎研究室との共同研究として行われました．江崎信芳教授，三原久明助手，村松久司氏，大崎元晴氏には，予測結果の検証として実際に実験を行って頂くとともに，論文作成時に必要な様々なデータを快く提供して頂きました．特に，三原久明助手には，本研究で開発した手法の応用面に関し，多くのディスカッションをして頂き，生化学の観点から多くの助言を頂きました．深く感謝致します．

また金久研究室およびバイオインフォマティクスセンターに関する全ての方々にも深く感謝の意を表したいと思います．阿久津研究室の方々，藤研究室の方々，馬見塚研究室の方々，金久研究室の秘書の方々，スーパーコンピューターラボラトリーの方々にもお世話になりました．

最後に，様々な面から研究生活および日常生活を支えて頂いた家族，友人の方々

に心より感謝いたします.

参考文献

[1] Akaho, S., A kernel method for canonical correlation analysis, *International Meeting of Psychometric Society (IMPS)*, 2001.

[2] Akutsu, T., Miyano, S., and Kuhara, S., Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function, *J. Comput. Biol.*, 7(3-4), 331-343, 2000.

[3] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215, 403-410, 1990.

[4] Anderson, T.W., An Introduction to multivariate statistical analysis, *Wiley, New York*, 1984.

[5] Asai, K., Hayamizu, S., and Handa, K., Prediction of protein secondary structure by the hidden Markov model, *Comput. Appl. Biosci.*, 9, 141-146, 1993.

[6] Bach, F.R. and Jordan, M.I., Kernel independent component analysis, *Journal of Machine Learning Research*, 3, 1-48, 2002.

[7] Bengio, Y., Vincent, P., Paiement, J.-F., Delalleau, O., Ouimet, M. and Le Roux, N., Spectral clustering and kernel PCA are learning eigenfunctions, *Technical Report 1239, Département d'informatique et recherche opérationnelle, Université de Montreal*, 2003.

[8] Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., Yuan, Y., Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem Sci*, 23, 324-328, 1998.

[9] Borser, B.E., Guyon, I.M., and Vapnik, V.N., A training algorithm for optimal margin classifiers, *Proceedings of the 5th annual ACM workshop on Computational Learning Theory*, 144-152, 1992.

[10] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., and Haussler, D., Knowledge-based analysis of microarray

gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA*, 97, 262-267, 2000.

[11] Cai, C.Z., Wang, W.L., Sun, L.Z., and Chen, Y.Z., Protein function classification via support vector machine approach, *Math. Biosci.*, 185, 2, 22-111, 2003.

[12] Chen, T., He, H.L., and Church, G.M., Modeling gene expression with differential equations, *Proc. Pac. Symp. on Biocomputing*, 29-40, 1999.

[13] Comon, P., Independent component anlaysis - a new concept?, *Signal Processing*, 36, 287-314, 1994.

[14] Eisen, M.B., Spellman, P.T., Patrick, O.B., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95, 14863-14868, 1998.

[15] Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A., Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402, 86-90., 1999

[16] Ermolaeva, M.D., White, O., and Salzberg, S.L., Prediction of operons in microbial genomes, *Nucleic Acids Res.*, 29, 1216–1221, 2001.

[17] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., Using Bayesian networks to analyze expression data, *J. Comput. Biol.*, 7(3-4), 601-620, 2000.

[18] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16, 906-914, 2000.

[19] Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., Cohen, F.E., Co-evolution of proteins with their interaction partners, *J. Mol. Biol.*, 299, 283-293, 2000.

[20] Gribskov, M. and Robinson, N.L., Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching, *Computers and Chemistry*, 20, 1, 25-33, 1996

[21] Hotelling, H., Relation between two sets of variates, *Biometrika*, 28, 322-377, 1936.

[22] Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K., Global analysis of protein localization in budding yeast, *Nature*, 425, 686-691, 2003.

[23] Hyvärinen, A., Fast and robust fixed-point algorithms for independent component analysis, *IEEE trans. on Neural Networks*, 10(3), 626-634, 1999.

[24] Hyvärinen, A., Survey on independent component analysis, *Neural Computing Surveys*, 2, 94-128, 1999.

[25] Ito, T., Takemoto, K., Mori, H., and Gojobori, T., Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes, *Journal of Molecular Evolution*, 16, 332–346, 1999.

[26] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. USA*, 98(8), 4569-4574, 2001.

[27] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M., A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, 302, 449-453, 2003

[28] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30, 42-46, 2002.

[29] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M., The KEGG resources for deciphering the genome, *Nucleic Acids Res.*, 32, D277-D280, 2004.

[30] Kondor, R.I. and Lafferty, J., Diffusion kernels on graphs and other discrete input, *Proc. Int. Conf. Machine Learning*, 315-322, 2002.

[31] Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D., Hidden Markov models in computational biology. Applications to protein modeling, *J. Mol. Biol.*, 235, 1501-1531, 1994.

[32] Lin, J. and Gerstein, M., Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels, *Genome Res.*, 10, 808-818, 2000.

[33] Liebermeister, W., Linear modes of gene expression determined by independent component analysis, *Bioinformatics*, 18(1), 51-60, 2002.

[34] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D., A combined algorithm for genome-wide prediction of protein function, *Nature*, 402, 83-86, 1999.

[35] Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B., An introduction to kernel-based keraning algorithms, *IEEE Trans. Neural Networks*, 12, 181-201, 2001.

[36] Ng, A.Y., Jordan, M.I., and Weiss, Y., On Spectral Clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems*, 14, 2001.

[37] Nakaya, A., Goto, S., and Kanehisa, M., Extraction of correlated gene clusters by multiple graph comparison, *Genome Informatics*, 44-53, 2001.

[38] Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M., A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Res.*, 28, 4021-4028, 2000.

[39] Park, K. and Kanehisa, M., Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, *Bioinformatics*, 19, 13, 1656-1663, 2003.

[40] Pavlidis, P., Weston J., Cai J., and Grundy, W.N., Gene functional classification from heterogeneous data, *RECOMB2001*, 249-255, 2001.

[41] Pazos, F. and Valencia, A., Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Engineering*, 14, 609-614, 2001.

[42] Pazos, F. and Valencia, A., In silico two-hybrid system for the selection of physically interacting protein pairs, *Proteins*, 47, 219-227, 2002.

[43] Pellegrini, M., Marcotte, E.M., Thompson, M. J., Eisenberg, D., and Yeates, T.O., Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA*, 96, 4285-4288, 1999.

[44] Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J., Operons in *Escheria coli*:Genomic analysis and prediction, *Proc. Natl. Acad. Sci. USA*, 97, 6652-6657, 2000.

[45] Schölkopf, B., Smola, A.J., and Müller, K.-R., Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, *Neural Computation*, 10, 1299-1319, 1998.

[46] Schölkopf, B., and Smola, A.J., Learning with Kernels, *MIT Press*, 2002.

[47] Schölkopf, B., Tsuda, K., and Vert, J.-P., Kernel Methods in Computational Biology, *MIT Press*, 2004.

[48] Smith, T.F. and Waterman, M.S., Identification of common modlecular subsequences, *J. Mol. Biol.*, 2004.

[49] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol. Biol. Cell.*, 9(12), 3273-3297, 1998.

[50] Tekaia, F., Lazcano, A., and Dujon, B., The genomic tree as revealed from whole proteome comparisons, *Genome Res.*, 9, 550-557, 1999.

[51] Toh, H. and Horimoto, K., Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics*, 18, 287-297, 2002.

[52] Tsuda, K., Kin, T., Asai, K., Marginalized kernels for biological sequences, *Bioinformatics*, 18, S268-S275, 2002.

[53] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, 10, 403(6770), 601-603, 2000.

[54] Vapnik, V.N., Statistical Learning Theory, *Wiley, New York*, 1998.

[55] Vert, J.-P., A tree kernel to analyze phylogenetic profiles, *Bioinformatics*, 18, S276-S284, 2002.

[56] Vert, J.-P. and Kanehisa, M., Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA, *Advances in Neural Information Processing Systems 15*, 1425-1432, MIT Press, Cambridge, MA, 2003.

[57] Vert, J.-P. and Kanehisa, M., Extracting active pathways from gene expression data, *Bioinformatics*, 19, 238ii-234ii, 2003.

[58] Walters, D.M., Russ, R., Knackmuss, H.J., and Rouviere, P.E., High-density sampling of a bacterial operon using mRNA differential display, *Gene*, 273, 2, 305-315, 2001.

[59] Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V., Genome trees and the tree of life, *Trends in Genetics*, 18, 9, 472-479, 2002.

[60] Wilbur, W.J. and Lipman, D.J., Rapid similarity searches of nucleic acid and protein data banks, *Proc. Natl. Acad. Sci. USA*, 80, 726-730, 1983.

[61] Yada, T., Nakao, M., Totoki, Y., and Nakai, K., Modeling and predicting transcriptional units of *Escheria coli* genes using hidden Markov models, *Bioinformatics*, 15, 987-993, 2001.

[62] Yamanishi, Y., Itoh, M. and Kanehisa, M., Extraction of organism groups from phylogenetic profiles using independent component analysis, *Genome Informatics*, 13, 61-70, 2002.

[63] Yamanishi, Y., Vert, J.-P., Nakaya, A. and Kanehisa, M., Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis, *Bioinformatics*, 19, i323-i330, 2003.

[64] Yamanishi, Y., Vert, J.-P., and Kanehisa, M., Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics*, 20, i363-i370, 2004.

[65] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M, Peng, V., Ngai, J., and Speed, T. P., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res.*, 15, 30, 4, e15, 2002.

[66] Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J. and Kasif, S., Computational identification of operons in microbial genomes, *Genome Res.*, 12, 8, 1221–1230, 2002.

[67] http://cib.nig.ac.jp/dda/taitoh/operondata.html

[68] http://mips.gsf.de/

[69] http://www.genome.jp/kegg/

[70] http://www.ncbi.nlm.nih.gov/

[71] http://yeastgfp.ucsf.edu/